# CRISPR activation screening at single-cell transcriptomic resolution: discovering new regulators of zygotic genome activation

**Celia Alda Catalinas**

Babraham Institute

Graduate School of Life Sciences

University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

Gonville and Caius College September 2019

# Declaration

I hereby declare that this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the Preface and specified in the text. The contents of this dissertation are original, except where specific reference is made to the work of others. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. This thesis contains less than 80,000 words excluding tables, bibliography and appendices, as granted by special permission from the Graduate School of Life Sciences of the University of Cambridge.

Celia Alda Catalinas

September 2019

# Abstract


## CRISPR activation screening at single-cell transcriptomic resolution: discovering new regulators of zygotic genome activation

### Celia Alda Catalinas

The maternal-to-zygotic transition in mammals consist of the passing of the developmental control from the mother to the embryo and involves two major events: the depletion of maternal products and the activation of the zygotic genome. Zygotic genome activation, or ZGA, coincides with dramatic epigenetic reprogramming processes and is likely orchestrated by the interplay of epigenetic and transcriptional regulators. In this dissertation, I develop a novel high-throughput screening method that combines pooled CRISPR activation (CRISPRa) with single-cell transcriptomics and apply it to discover new epigenetic and transcriptional regulators of ZGA in mouse embryos.

In chapter 1, I introduce the biological context and latest technological advances related to CRISPR and single-cell transcriptomics. In chapter 2, I outline the materials and methods used in this dissertation. Chapters 3, 4 and 5 are a description and discussion of the results obtained. First, in chapter 3, I optimise the tools necessary to perform a CRISPRa screen with single-cell RNA-sequencing read-out. Specifically, I test CRISPRa in mouse embryonic stem cells (mESCs) using two well characterised markers of ZGA as targets and, subsequently, I develop and validate a method for detecting short guide RNA (sgRNA) molecules in single-cell RNA sequencing libraries.

In chapter 4, I select 230 screen candidates and construct a lentiviral sgRNA library to target their promoters by CRISPRa. After transducing this library into mESCs and performing single-cell RNA-sequencing for approximately 300,000 cells, I assess target gene activation by CRISPRa for each candidate gene and investigate different sgRNA features as well as several

transcriptional and epigenetic parameters in relation to CRISPRa efficiency at the single-cell level.

In chapter 5, I discover 44 screen hits using unsupervised methods to identify latent sources of gene expression variation and differential gene expression analysis. Amongst the top hits, I identify the DNA binding protein *Dppa2*, the chromatin remodeller *Smarca5* and the transcription factor *Patz1*. I subsequently validate these top hits using both CRISPRa and cDNA overexpression in mESCs. By doing this, I uncover new factors that promote a ZGA-like response in mESCs and are, therefore, strong candidates for ZGA regulation *in vivo*. Further molecular characterisation of *Dppa2* and *Smarca5* reveals that *Smarca5* induces a ZGA-like transcriptional signature in mESCs via *Dppa2*, which in turns acts through the zygotic transcription factor *Dux*.

Lastly, in chapter 6, I summarise my findings, discuss their relevance in light of the literature in the field and propose future directions that arise from my conclusions.

Altogether, by developing and applying a new method for single-cell transcriptomic profiling of CRISPRa-perturbed cells, I provide novel system-level insights into the molecular mechanisms that orchestrate ZGA.

# Acknowledgements

This thesis, and my whole PhD journey during the last four years, would not have been possible without the support of many people that I would like to acknowledge. First and foremost, I would like to thank my supervisor, Wolf Reik, for giving me the opportunity to work in his group, for transmitting me his enthusiasm in epigenetics and for his trust and guidance while letting me conduct my independent research.

I am most grateful to Melanie Eckersley-Maslin for being not only my scientific mentor but also a constant support during my PhD. Thank you for introducing me in the exciting field of early development, for your guidance not only scientifically but also personally, for being an example of hard work and commitment while enjoying the little things in life, for always trying to get the most out of me and for supporting me both in the happy and toughest moments. This PhD would not have been the same without you.

I have been really lucky to work and collaborate with many other amazing scientists. I would like to thank Irene Hernando-Herraez for her help and advice but, more importantly, for being a great friend in and outside the lab. My special gratitude goes also to Fátima Santos for her support, advice and discussions throughout my time at the Babraham Institute. Also, a special mention to Stephen Clark, for his initial training in single-cell library preparation. Thank you to Wendy Dean for her support, ideas and suggestions. I am also grateful to Christel Krueger, Felix Krueger and Simon Andrews for their bioinformatics advice. Thanks to Oana Kubinyecz for sharing your enthusiasm in the last year. Thank you as well to Peter Rugg-Gunn for being my PhD advisor. I also need to extend my gratitude to my collaborators in this project, Danila Bredikin and Oliver Stegle, for your contribution and helpful insights. And last but not least, thank you to all past and present members of the Reik lab for helpful discussions and for contributing to make such a stimulating work environment.

This PhD research would not have been possible without the funding from BBSRC DTP.

These last four years in Cambridge would not have been the same without my friends who I would like to thank. To Julia, my friend and housemate, for being there from the beginning, and even earlier, for the fun times but also for always being by my side. To Vicen and Irene, you know how much you mean to me. To my friends Nacho, Alba, Hassal and Chiara for making this a fun experience. To my PhD lab mates, Julia, Poppy, Rebecca, for sharing this journey with me. To my colleagues in the Epigenetics Department, for the fun at the bar on Fridays. Being away from home is not easy, so I am deeply grateful to my friends that have been supporting me in the distance and welcoming me whenever I went home. To Jose, for your friendship and support no matter what. To Maria, Nelson, Bea, Nico, Fran, Dani and Lore, for making me feel as if I never left. To Alberto and Lorena, for the occasional meet-ups and making me remember the incredible times we shared in Uppsala. To Bea, Laura, Alba, Isa, Lorena and Ixchel, for making the most out of my visits to Sigüenza, sharing time with you is always great fun. And to Araceli, Marta and Ricardo, you have been great friends despite the distance and our laughter, even by texts, always makes things easier.

And lastly, I switch to Spanish to thank my family. Papá, mamá, Paula, esto nunca lo habría conseguido sin vosotros. Gracias por creer siempre en mí y apoyarme, por hacerme fuerte en los momentos más difíciles pero también por compartir las alegrias de mis éxitos. Gracias por inculcar en mí que todo es posible trabajando duro pero también hacerme ver que la vida hay que disfrutarla. En definitiva, gracias a los tres por convertirme en la persona que soy hoy.

# Table of Acknowledgement of Assistance

**Initial training in techniques and laboratory practice and subsequent mentoring:**

Melanie Eckersley-Maslin provided initial training in cell culture, molecular biology techniques and next-generation sequencing analysis. She also provided continuous mentoring throughout my PhD.

**Data / materials obtained from a technical service provider:**

- Buk RNA-sequencing libraries (sections 3.2.1, 3.2.2 and 5.2.3) were constructed and sequenced by the Wellcome Trust Sanger Institute bespoke pipelines.
- Single-cell RNA-sequencing libraries in section 3.2.3 were sequenced at the the Wellcome Trust Sanger Institute.
- Single-cell RNA-sequencing libraries in section 3.2.4 and for the screen (chapters 4 and 5) were constructed and sequenced by the CRUK Genomics facility.
- Single-cell NMT-seq libraries (in section 4.2.4) were sequenced at the Sequencing facility of the Babraham Institute.
- Next-generation sequencing data was mapped by the Bioinformatics facility at the Babraham Institute.
- FACS was performed by the Flow cytometry facility at the Babraham Institute.
- The oligo sgRNA library was synthesised by Twist Bioscience. It was then cloned into the CROP-sgRNA-MS2 backbone and packaged into a lentiviral library by VectorBuilder.

**Data / analysis produced jointly:**

- The scRNA-seq data presented in section 3.2.3 was analysed jointly with Irene Hernando-Herraez (Reik laboratory).
- Fátima Santos (Reik laboratory) helped with image acquisition of DPPA3 immunofluorescence stainings (Figure 3.30).

- The variability analysis presented in Figure 4.29 was done jointly with Irene Hernando-Herraez (Reik laboratory).
- Christel Krueger (Bioinformatics facility, Babraham Institute) assisted with scNMT-seq data analysis (in section 4.2.4).
- Processing of the screen data (section 4.2.3), implementation of the MOFA model and screen hit calling (sections 5.2.1 and 5.2.2) was performed in collaboration with Danila Bredikin and Oliver Stegle (EMBL, Heidelberg).
- Laura Benson (Reik laboratory) provided technical help for the qPCRs whose results are presented in Figures 5.29-5.32.

**Data / materials provided by someone else:**

- The immunofluorescence stainings for DPPA2 and SMARCA5 in early embryos (Figures 5.27 and 5.28) were done by Fátima Santos and Oana Kubinyecz (Reik laboratory).
- *Smarca5* KO mESCs were a kind gift from Dirk Schübeler (FMI, Basel).
- *Dppa2* KO mESCs were generated by Melanie Eckersley-Maslin (Reik laboratory).
- The following plasmids were obtained from Addgene as a gift from the following laboratories:
    - From Didier Trono (EPFL, Lausanne): pMD2.G (Addgene, 12259) and psPAX2 (Addgene, 12260).
    - From Feng Zhang (Broad Institute, Cambridge, MA): lenti dCas9-VP64_Blast (Addgene, 61425), lenti MS2-p65-HSF1_Hygro (Addgene, 61426) and lenti sgRNA(MS2)_puro backbone (Addgene, 73795).

# Publications

**Alda-Catalinas, C**., Bredikhin, D., Hernando-Herraez, I., Kubinyecz, O., Santos, F., Eckersley-Maslin, M.A., Stegle, O., Reik, W., 2019. A single-cell transcriptomics CRISPR-activation screen identifies new epigenetic regulators of zygotic genome activation. *bioRxiv*, p.741371.

Eckersley-Maslin, M.A.*, **Alda-Catalinas, C***. & Reik, W., 2018. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nature Reviews Molecular Cell Biology*, 19(7), pp.436–450.

* equal contribution

Eckersley-Maslin, M., **Alda-Catalinas, C**., Blotenburg, M., Kreibich, E., Krueger, C. & Reik, W., 2019. Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes & Development*, 33(3-4), pp.194-208

Clark, S.J., Argelaguet, R., Kapourani, C.A., Stubbs, T.M., Lee, H.J., **Alda-Catalinas, C.**, Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C., Stegle, O., Reik, W., 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9(1), p.781.

# List of Abbreviations and Acronyms

| | |
|---|---|
| ATAC-seq | Assay for Transposable Accessible Chromatin followed by sequencing |
| BER | Base-Excision Repair |
| Bp | Base pairs |
| BSA | Bovine Serum Albumin |
| CAGE | Cap Analysis of Gene Expression |
| ChIP-seq | Chromatin Immunoprecipitation followed by sequencing |
| CRISPR | Clustered Regularly Interspaced Short Palindromic Repeat |
| CRISPRa | CRISPR-activation |
| dCas9 | dead Cas9 |
| DEGs | Differentially Expressed Genes |
| DNase-seq | DNase I hypersensitive sites sequencing |
| EF1α | Elongation Factor-1 alpha |
| ERV | Endogenous Retrovirus |
| ESCs | Embryonic Stem Cells |
| FACS | Fluorescence-Activated Cell Sorting |
| FBS | Fetal Bovine Serum |
| FDR | False Discovery Rate |
| FRAP | Fluorescent Recovery After Photobleaching |
| GEMs | Gel beads-in-Emulsion |
| GO | Gene Ontology |

| | |
|---|---|
| GV | Germinal Vesicle |
| GVBD | Germinal Vesicle Breakdown |
| HIV | Human Immunodeficiency Virus |
| IRES | Internal Ribosome Entry Site |
| IVF | *In Vitro* Fertilisation |
| lncRNAs | Long non-coding RNAs |
| LTR | Long Terminal Repeats |
| MEGs | Maternal Effect Genes |
| mESCs | Mouse Embryonic Stem Cells |
| MII | Metaphase II |
| miRNA | Micro RNA |
| MOFA | Multi-Omics Factor Analysis |
| MOI | Multiplicity-Of-Infection |
| mRNA | Messenger RNA |
| MZT | Maternal-to-Zygotic Transition |
| NGS | Next-Generation Sequencing |
| NPBs | Nucleolar Precursor Bodies |
| ORF | Open Reading Frame |
| PAM | Protospacer Adjacent Motif |
| PBS | Phosphate-Buffered Saline |
| PCA | Principal Component Analysis |
| PCR | Polymerase Chain Reaction |
| PFA | Paraformaldehyde |
| PGCs | Primordial Germ Cells |
| Pol II | Polymerase II |
| Pol III | Polymerase III |
| PTMs | Post-Translational Modifications |

| | |
|---|---|
| qPCR | Quantitative reverse transcription PCR |
| RPKM | Reads Per Kilobase per Million |
| RPM | Reads Per Million |
| SAM | Synergistic Activator Mediator |
| SCNT | Somatic Cell Nuclear Transfer |
| scRNA-seq | Single-cell RNA-sequencing |
| scNMT-seq | Single-cell NOMe, methylation and transcription-sequencing |
| sgRNA | Synthetic guide RNA |
| siRNA | Small interfering RNA |
| TES | Transcription End Sites |
| TSS | Transcription Start Sites |
| UMI | Unique Molecular Identifier |
| UTR | Untranslated Region |
| ZGA | Zygotic Genome Activation |

# Contents

# List of Figures

# List of Tables

# List of Appendices

# Chapter 1 Introduction

Embryonic development in mammals starts with the fertilisation of an oocyte with a sperm, giving rise to a one-cell embryo, also called zygote. One of the most intriguing and studied aspects of development is how this unique cell is able to give rise to the plethora of cell types that constitute a fully-developed organism. Cell-fate decision rely on epigenetic modifications and activation of transcriptional programmes that determine cell function (reviewed in Allis & Jenuwein 2016). In development, the first transcriptional event is called zygotic genome activation (ZGA), and it is a tightly orchestrated and conserved process that is essential for continued development (reviewed in Vastenhouw et al. 2019; Eckersley-Maslin, Alda-Catalinas & Reik 2018). ZGA occurs as part of a wider developmental transition, the maternal-to-zygotic transition (MZT), in which developmental control is transferred from the mother to the embryo. Epigenetic reprogramming occurs hand-in-hand with transcriptional changes during the MZT and leads to a change in cellular identities that is essential for subsequent development (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018; Jansz & Torres-Padilla 2019; Ladstatter & Tachibana 2019). Understanding the molecular mechanisms that regulate ZGA in the context of the MZT means understanding the first transcriptional event in a living organism and therefore constitutes essential knowledge for basic reproductive biology and cellular reprogramming, but it is also highly relevant for clinical applications such as assisted reproduction and regenerative medicine.

In this dissertation, I developed a novel screening method using state-of-the-art technologies, such as CRISPR and single-cell RNA-sequencing, to uncover novel maternal regulators of mouse ZGA. In this chapter, I start by introducing the biological context of ZGA, focusing primarily on mouse and human and in the epigenetic remodeling processes that occur as part of the MZT. Next, I explain *in vitro* alternatives to study ZGA in mouse embryonic stem cells. In the second half of this introduction, I review the latest technological advances in single-cell sequencing, CRISPR, and the power of combining both.

## 1.1 The maternal-to-zygotic transition

In mammalian embryos, after fertilisation, the very first cellular decisions are made by maternally-provided RNA and proteins stored in the cytoplasm of the oocyte and deposited into the zygote (reviewed in Lei Li et al. 2013; Schultz 2002; Vastenhouw et al. 2019). The MZT involves the depletion of this maternal deposits (reviewed in Walser & Lipshitz 2011), activation of the zygotic genome (ZGA) (reviewed in M. T. Lee et al. 2014; Jukam et al. 2017) and extensive epigenetic remodeling processes (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018; Jansz & Torres-Padilla 2019; Ladstatter & Tachibana 2019). Together, these events dramatically remodel the embryonic gene expression landscape and cellular identities, and are highly coordinated to ensure a correct developmental progression.

### 1.1.1 From oocytes to blastocysts

#### 1.1.1.1 Oocyte growth and meiotic maturation

Before fertilisation, oocytes need to acquire meiotic and developmental competence. This process goes back to mouse embryonic development day 7.25 (E7.25), when primordial germ cells (PGCs) are first detected (Ginsburg et al. 1990). In the latter stages of fetal development, at E13.5, PGCs enter meiosis and arrest at the end of prophase I (Figure 1.1). At this stage, oocytes are contained within primordial follicles (Figure 1.1), where the surrounding somatic cells – the granulosa cells – establish critical signalling connections for oocyte growth (reviewed in Rong Li & Albertini 2013). After birth, the pool of primordial follicles constitutes the entire store of germ cells available for reproduction and, periodically, a cohort of primordial follicles enters into a growth phase (reviewed in Amleh & Dean 2002). During growth, oocytes suffer dramatic gene expression changes to ensure storage of organelles and macromolecules that will be needed for early stages of embryo development (Pan et al. 2005). Concomitantly, granulosa cells proliferate to eventually give rise to pre- or early antral follicles (reviewed in Rong Li & Albertini 2013), which grow into antral follicles after puberty with the release of follicle-stimulating hormone (FSH) (Kumar et al. 1997) (Figure 1.1).

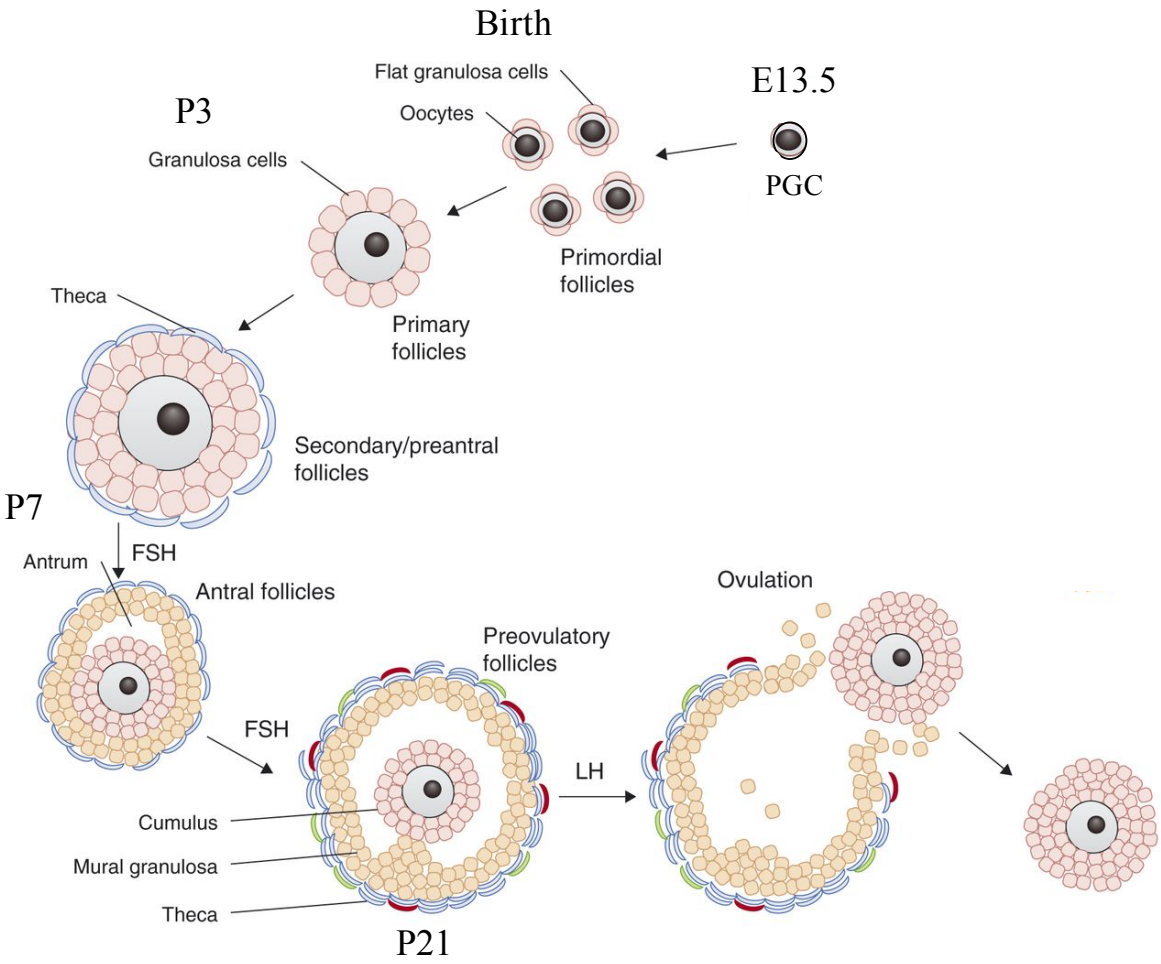



**Figure 1.1– Schematic representation of mouse follicle development**

Mouse follicles develop from primordial germ cells (PGC), which enter meiosis at embryonic day 13.5 (E13.5). At birth, the pool of primordial follicles constitutes the entire store of germ cells available for reproduction and, in these structures, oocytes are surrounded by epithelial cells called granulosa cells. Primordial follicles develop first into primary follicles at three days post-birth (P3) and into secondary and antral follicles at P7, when they become surrounded by stromal cells called theca cells. In antral follicles, granulosa cells are differentiated into cumulus cells that surround the oocyte and mural granulosa cells that surround the antrum. The development from secondary to antral follicles is induced by follicle-stimulating hormone (FSH), which further drives development of antral follicles into preovulatory follicles at P21. Lastly, the luteinizing hormone (LH) surge induces meiotic maturation and ovulation of metaphase II (MII) oocytes. Image adapted from Georges et al. 2014.

Approximately every 4 to 6 days in mouse and 28 days in human, before ovulation, the luteinizing hormone (LH) surge induces fully grown germinal vesicle (GV) oocytes to complete the first meiotic division, extruding the first polar body, and enter into meiosis II (reviewed in Russell & Robker 2007). During this process, oocytes acquire meiotic maturation competence in two steps: firstly, by undergoing germinal vesicle breakdown (GVBD) to exit prophase I and progress through metaphase I, and secondly, by starting their second meiotic division, where they get arrested in metaphase II (MII) (Sorensen & Wassarman 1976; Wickramasinghe et al. 1991). Finally, MII oocytes are released from the ovaries during ovulation (Figure 1.1), ready to be fertilised by a sperm.

#### 1.1.1.2 Fertilisation and egg activation

With fertilisation, oocytes acquire developmental competence. Fertilisation is defined as the fusion of the male and female gametes and it involves a series of coordinated, sequential molecular events. Following ovulation, the MII oocyte contains an extracellular matrix called zona pellucida that is composed of three sulphated glycoproteins, ZP1, ZP2 and ZP3 (reviewed in J. P. Evans & Florman 2002). This structure mediates the initial oocyte-sperm recognition and is responsible for triggering the sperm acrosome reaction (reviewed in Georgadaki et al. 2016). Specifically, after capacitation, the sperm recognises and binds primarily to ZP3 glycoproteins via O-linked oligosaccharide connections (Bleil & Wassarman 1980; Florman & Wassarman 1985). This recognition triggers signal transduction events in the sperm that include the opening of low voltage-activated calcium channels (Arnoult et al. 1996), which eventually results in a sustained calcium influx that drives exocytosis of the acrosome, a single secretory granule present in the head of mammalian sperm that is required for fertilisation (reviewed in J. P. Evans & Florman 2002). The acrosome reaction allows sperm attachment to the oocyte membrane, cell-cell adhesion and membrane fusion of the two gametes mediated by multimeric complexes (reviewed in J. P. Evans 1999; Georgadaki et al. 2016). Once both gametes have fused, low-frequency oscillations in cytosolic calcium concentration from sequestered calcium in the endoplasmic reticulum of the fertilised oocyte induce four events that allow the commencement of embryonic development: 1) meiotic resumption with extrusion of the second polar body, 2) release of cortical granules that modify the zona

pellucida and prevent polyspermy, 3) formation of female and male pronuclei, and 4) the first cleavage division of the embryo (reviewed in Schultz & Kopf 1995; Krauchunas & Wolfner 2013).

#### 1.1.1.3 Pre-implantation development. Totipotency and pluripotency

In mice, pre-implantation development lasts four and a half days, spanning from fertilisation to implantation into the mother's uterus (reviewed in Wennekamp et al. 2013). After fertilisation, the egg completes meiosis and a one-cell embryo or zygote is formed, containing haploid paternal and maternal pronuclei. Then, the DNA is replicated and the two pronuclei undergo syngamy to form a diploid one-cell embryo (Zamboni et al. 1972), which enters a series of cleavage division that result first in the formation of a two-cell embryo after the first mitosis, and eventually, in the formation of a structure called blastocyst. While these cleavage divisions occur, two essential processes take place: firstly, the handover of the developmental control from the mother to the embryo, which involves the depletion of maternal products and the transcriptional activation of the zygotic genome, and secondly, the first two lineage decisions in development, where cells commit to the embryo proper or to extra-embryonic tissues that are going to give rise to the placenta (reviewed in Leung & Zernicka-Goetz 2015) (Figure 1.2).

During the initial stages of pre-implantation development, the embryo is totipotent, that is, all blastomeres or cells that constitute it have the potential to give rise to any cell type, both embryonic and extra-embryonic (reviewed in Ishiuchi & Torres-Padilla 2013; Leung & Zernicka-Goetz 2015; Zhou & Dean 2015; Dang-Nguyen & Torres-Padilla 2015) (Figure 1.2). The exact developmental stage until which the embryo retains this totipotent potential is controversial. The most stringent test of totipotency, proposed by Tarkowski in 1959, consists on isolating a single blastomere, placing it into an empty zona pellucida, returning this construct to the oviduct and monitor developmental progression (TARKOWSKI 1959). These early experiments defined totipotency up to the two-cell stage in the mouse embryo. However, later on, it was suggested that the failure of four-cell and eight-cell embryo blastomeres to perform successfully in this test was due to the limited size of the inner cell mass (ICM) formed rather than a restricted developmental potential of cells at these stages (Rossant 1976). In this

respect, individual blastomeres from 16-cell stage embryos supported by carrier labelled tetraploid blastomeres were able to contribute to embryonic and extra-embryonic lineages (TARKOWSKI et al. 2010). More recently, a study examining the developmental potential of the blastomeres in two-cell embryos suggested an unequal totipotent potential of the two blastomeres (Casser et al. 2017), arising from transcriptomic differences between them (Casser et al. 2018). In human embryos, experiments where individual blastomeres were split at the four-cell stage and individually cultured *in vitro* up to 6 days showed that each blastomere retained the capacity to form a blastocyst containing a trophectoderm and an ICM (van de Velde et al. 2008). However, due to the technical manipulations done to embryos in all these studies aiming to define totipotency, its definition remains unclear and continuously under revision (Boiani et al. 2019).

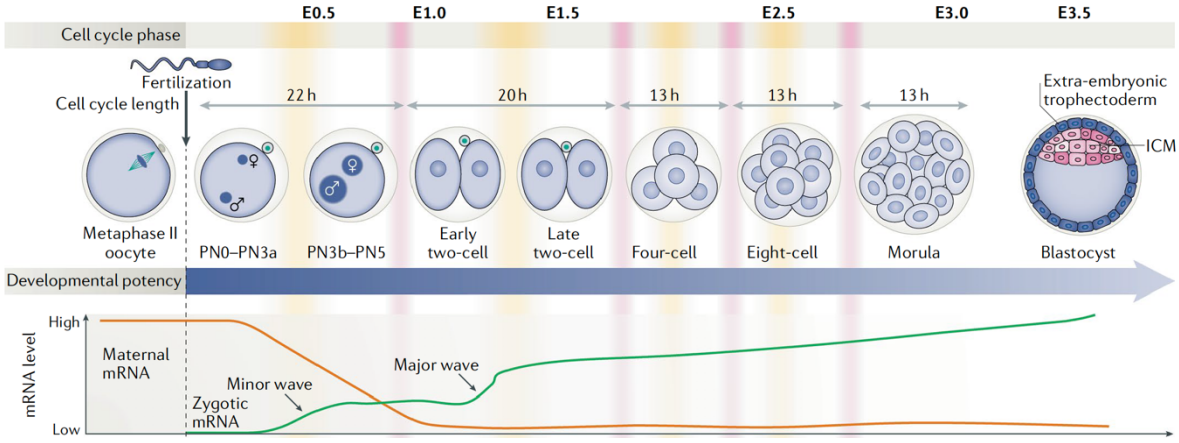



**Figure 1.2– Overview of cleavage divisions and transcriptional changes during pre-implantation development**

Following fertilisation, a series of cleavage divisions from the mouse zygote through to the blastocyst take place, accompanied by a decline in potency. The early blastocyst is formed by inner cell mass cells (ICM) that will give rise to the embryo proper, and extra-embryonic trophectoderm cells, that will give rise to the placenta. Maternal mRNA stores (orange line) are depleted during the first hours of development while two waves of transcriptional zygotic genome activation (ZGA; green line) occur: the first or minor wave predominantly in the male pronucleus (PN) in the zygote, and the second or major wave in mid-two-cell to late-two-cell embryos. Approximate cell cycle phases are represented by yellow bars for S phase and red bars for M phase. Approximate embryonic days (E) are denoted. Image adapted from Eckersley-Maslin, Alda-Catalinas & Reik 2018.

With the loss of potency, the first lineages in development are specified at E3.5, when cells are segregate into the ICM and the trophectoderm (Figure 1.2), as specified from asymmetric cell divisions from the four-cell stage (Johnson & Ziomek 1981; Sutherland et al. 1990; Piotrowska-Nitsche et al. 2005). The cells in the ICM continue differentiating into the epiblast- which will constitute the embryo proper-, and the primitive endoderm -another extraembryonic tissue-, which are set apart by E4.5 (reviewed in Rossant & Tam 2009). The pre-implantation epiblast in the blastocyst has a more restricted developmental potential as it can only give rise to embryonic, but not to extraembryonic, tissues; this feature is defined as pluripotency (reviewed in De Los Angeles et al. 2015).

### 1.1.2 Regulation of maternal gene products

After fertilisation, the embryo is transcriptionally quiescent and development relies on proteins and RNAs provided from the egg cytoplasm (reviewed in Wassarman & Kinloch 1992). While the stores deposited from the sperm were traditionally thought to have minimal contribution to early embryogenesis (reviewed in Sutovsky & Schatten 2000), recent studies are beginning to appreciate the importance of sperm components, such as tRNAs, during early events such as ZGA (Sharma et al. 2016). Nevertheless, for the purpose of this dissertation, and given the central role of the oocyte deposits for survival of the embryo after fertilisation, I will focus on the maternal regulation of early development.

Approximately 12 hours before ovulation, mouse oocytes become transcriptionally quiescent (Clegg & Piko 1983; De Leon et al. 1983). However, before that, as the oocyte is growing while arrested in prophase I, it transcribes and accumulates transcripts, most of which are immediately translated into proteins that will be deposited into the embryo (De Leon et al. 1983; Vitale et al. 2007). Other transcripts remain dormant and only become activated later during oocyte maturation by cytoplasmic polyadenylation (Racki & Richter 2006). Cytoplasmic polyadenylation of dormant transcripts leads to the translation of crucial proteins for meiotic maturation and early stages of development, such as the activation of the embryonic genome or ZGA, mitotic progression of the embryo throughout cleavage divisions in pre-implantation development, and establishment of the first lineage decisions. These factors are

known as maternal-effect genes (reviewed in Lei Li et al. 2010) and are further reviewed in section 1.1.6.

In mouse early embryos, most deposited maternal transcripts are degraded by the two-cell stage (Clegg & Piko 1983) (Figure 1.2), although their corresponding proteins can persist for much longer and have functions during the first lineage allocation (Nothias et al. 1996; Y. Gao et al. 2017). Although some maternal transcripts are degraded during meiotic maturation, the rate is slow (Clegg & Piko 1983; De Leon et al. 1983). Fertilisation triggers maternal messenger RNA (mRNA) destabilisation of transcripts that encode proteins with roles in specific events during meiosis, such as Growth Differentiation Factor 9 (GDF9) (Alizadeh et al. 2005; Su et al. 2007). Other transcripts are degraded in a zygotically-directed wave of degradation that follows zygotic genome activation after the two-cell stage (Hamatani et al. 2004; reviewed in Svoboda et al. 2015). Although less well studied in human embryos, gene expression profiling also suggests two waves of maternal transcript clearance (Dobson et al. 2004; P. Zhang et al. 2009; Xue et al. 2013).

As proposed by Vastenhouw *et al*. in a recent review, this coordinated regulation of maternal transcript clearance suggests tight regulation of transcript stabilisation and destabilisation mediated by RNA-binding proteins, small non-coding RNAs, RNA modifications and "codon optimality" (reviewed in Vastenhouw et al. 2019). The Y-box-binding protein MSY2 was found to stabilize maternal transcripts after being phosphorylated by the Cell Division Control protein 2 homolog A (CDC2A) kinase upon oocyte maturation (Medvedev et al. 2008; Medvedev et al. 2011; J. Yu et al. 2001; J. Yu et al. 2002). The best studied RNA binding proteins involved in mRNA destabilisation are the CCCH tandem zing finger protein 36-like 2 (ZFP36L2) (Ramos et al. 2004) and the B-cell translocating gene 4 (BTG4) (C. Yu, Ji, Sha, et al. 2016; Yusheng Liu et al. 2016). Disruption of either of these proteins in mouse oocytes impairs development after the one- to -two-cell stage (Ramos et al. 2004; C. Yu, Ji, Sha, et al. 2016; Yusheng Liu et al. 2016). Small non-coding RNAs play a role in transcript degradation during meiotic maturation, specifically, microRNAs (miRNAs) and small interfering RNAs (siRNAs) which act in an RNA-induced silencing complex (RISC) that contains Argonaute proteins (Murchison et al. 2007; Tang et al. 2007; Kaneda et al. 2009; Lykke-Andersen et al. 2008). In recent years, it has been discovered that chemical modifications of mRNAs can

directly influence post-transcriptional gene regulation (reviewed in Lewis et al. 2017). For example, the maternal protein YTH $N^6$-methyladenosine RNA binding protein 2 (YTHDF2), which binds $N^6$-methylation of adenosine (m6A) in mRNA, and the mRNA deadenylation protein CCR4-NOT transcription complex subunit 6-like (CNOT6L) induce destabilisation of mRNAs (Ivanova et al. 2017; Sha et al. 2018). Another form of mRNA modification seen to play a role during the MZT in mouse is terminal uridylation in deadenylated mRNAs by the terminal uridylyl transferases 4 and 7 (TUT4 and TUT7, respectively) (Morgan et al. 2017; Chang et al. 2018). Lastly, in the zebrafish MZT, it was found that transcripts enriched for non-optimal codons (a codon encoding an amino acid for which an alternative codon has higher cognate tRNA abundance) are preferentially degraded (Bazzini et al. 2016; Mishima & Tomari 2016), a regulation mechanism that has been proposed to be conserved in mice (Bazzini et al. 2016).

Importantly, not only RNAs but also maternal proteins are degraded after fertilisation and throughout pre-implantation development. This is mediated by ubiquitin-proteasome pathways with factors such as Ret Finger Protein-Like 4 (RFPl4) (Suzumori et al. 2003), Ubiquitin-conjugating enzyme E2 A (UBE2A) (Roest et al. 2004) and E3 ubiquitin-protein ligase RNF114 (Y. Yang et al. 2017), and by autophagy (TSUKAMOTO et al. 2008).

### 1.1.3 Transcriptional dynamics during zygotic genome activation

In mouse and human embryos, the initially quiescent zygotic genome becomes activated in two transcriptional bursts, namely, the minor wave and the major wave of ZGA (reviewed in Schultz 2002; Lei Li et al. 2013; M. T. Lee et al. 2014; Jukam et al. 2017; Eckersley-Maslin, Alda-Catalinas & Reik 2018; Vastenhouw et al. 2019) (Figure 1.2). In total, between the two waves, approximately 20% of the coding genome is transcribed in mouse during the early days of development (Hamatani et al. 2004).

#### 1.1.3.1 Minor wave of zygotic genome activation

The first or minor wave occurs predominantly in the male pronucleus of the mouse zygote, in the S to G2 cell cycle phases (Figure 1.2) around 10 hours post-fertilisation, and it involves

low-level, genome-wide and promiscuous transcription (Ram & Schultz 1993; Bouniol et al. 1995; Majumder & DePamphilis 1995; Nothias et al. 1996; Aoki et al. 1997; F. Zeng et al. 2004; Xue et al. 2013; Deng et al. 2014; Abe et al. 2015; S.-J. Park et al. 2015). The number of transcripts expressed during this wave has been analysed by single-cell RNA-sequencing and varies between approximately 500 and 4000 coding genes, depending on the study (Xue et al. 2013; Abe et al. 2015; S.-J. Park et al. 2015; Deng et al. 2014), likely due to slight differences in the exact developmental stage analysed, the type of RNA libraries (total RNA or poly(A)-capture) and the sequencing depth.

A few of the transcripts produced during minor ZGA are only transiently transcribed, but the majority continue being transcribed in two-cell embryos, during the major wave of ZGA (S.-J. Park et al. 2015; Abe et al. 2015). Nevertheless, given the pervasive nature of transcription during minor ZGA, it is not only coding transcripts but also intergenic regions, including transposable elements (see section 1.1.3.3), that are transcribed even in the absence of core-promoter elements (Abe et al. 2015) or enhancers (Wiekowski et al. 1991; Majumder & DePamphilis 1995), requiring only minimal promoter features (Hamamoto et al. 2014). Moreover, it has been proposed that despite this pervasive nature, there is some selectivity towards sperm-primed transcripts since more than 300 minor ZGA transcripts are also expressed in sperm but not in parthenogenetic embryos (embryos originated from an activated MII oocyte in the absence of fertilisation from sperm) (S.-J. Park et al. 2015).

The function of transcription during minor ZGA is mostly unknown, although it is believed that most of the mRNAs produced are not functional due to inefficient 3' end processing and splicing (Abe et al. 2018), and transcription-translation uncoupling in plasmid-borne reporter genes in one-cell embryos (Nothias et al. 1996). Importantly, even though the male pronucleus supports most of the transcription detected in one-cell embryos (Henery et al. 1995; Aoki et al. 1997), likely due to the drastic chromatin remodeling that it suffers after fertilisation (see section 1.1.4), unspliced transcripts were also detected in parthenogenetic one-cell embryos (S.-J. Park et al. 2015; Abe et al. 2015), indicating that ineffective splicing is not exclusive to paternal transcripts. Despite the uncertainty in the role of transcription in the one-cell embryo, being unclear whether it has a specific function or it is merely a consequence of a relaxed chromatin structure (see section 1.1.4), transcription inhibition experiments have effectively

shown that promiscuous transcription needs to occur in the embryo before the more regulated wave of major ZGA takes place (Abe et al. 2018).

Although less well characterised, low transcriptional rates occur from the zygote to the four-cell stage in human embryos, equivalent to a minor wave of ZGA (Xue et al. 2013; L. Yan et al. 2013). However, clustering analysis using transcriptomic information show that the one-cell embryo is transcriptionally distinct from later cleavage stages and from the oocyte, suggesting a unique transcriptomic landscape (Xue et al. 2013; L. Yan et al. 2013) that probably reflects the presence of maternal products and new zygotic transcription.

#### 1.1.3.2 Major wave of zygotic genome activation

The major wave of ZGA occurs at the S to G2 cell-cycle phase of the two-cell stage in mouse embryos, after the first cleavage division around 19 hours post fertilisation (Majumder & DePamphilis 1995; Aoki et al. 1997; Hamatani et al. 2004; F. Zeng et al. 2004; Xue et al. 2013; Deng et al. 2014; S.-J. Park et al. 2015) (Figure 1.2), while it is not until the eight-cell stage that a major wave of transcription is detected in the human embryo (Braude et al. 1988; Dobson et al. 2004; Vassena et al. 2011; Xue et al. 2013; L. Yan et al. 2013). The major wave of ZGA involves a shift in cellular identity from oocyte-like to embryo-like, with the degradation of maternal mRNAs and accumulation of novel zygotic transcripts that convey cellular totipotency (reviewed in Svoboda et al. 2015; Baker & Pera 2018).

Similar to minor ZGA, the number of transcripts reported to be involved in major ZGA depends on the technology used to detect them (microarray or next-generation sequencing), but, regardless, it is known to involve thousands of genes (F. Zeng et al. 2004; F. Zeng & Schultz 2005; S.-J. Park et al. 2015; Deng et al. 2014; Xue et al. 2013). While most transcripts from minor ZGA continue to be transcribed, this wave of transcription in mouse two-cell embryos and human eight-cell embryos differs from minor ZGA not only in the volume but also in the transcripts produced (Xue et al. 2013; S.-J. Park et al. 2015; Abe et al. 2018). Although most of the genes transcribed during major ZGA have unknown functions, this wave also involves expression of genes related to transcription regulation, RNA processing, ribosome biogenesis and protein translation which are key for later pre-implantation stages (Xue et al. 2013).

Furthermore, single-cell sequencing of hybrid zygotes showed the existence of stochastic monoallelic expression (Deng et al. 2014).

Development to the two-cell stage and major ZGA is accompanied by genome-wide accumulation of repressive histone marks (Santos et al. 2005). This leads to a transcriptionally repressive state (Majumder et al. 1993; Wiekowski et al. 1993) in a transition that is dependent on DNA replication (Wiekowski et al. 1991; Majumder et al. 1993; Henery et al. 1995; Forlani et al. 1998). Consistently, many genes with unknown functions are only transiently expressed during major ZGA (Davis et al. 1996; Xue et al. 2013; S.-J. Park et al. 2015) and their expression is enhancer-dependent (Wiekowski et al. 1991; Majumder et al. 1993; Henery et al. 1995).

#### 1.1.3.3 Transposable and repeat elements during zygotic genome activation

Transposable elements are mobile and repetitive genomic regions that are able to change their location within the genome by transposition. In mammals, they comprise about 70% of the genome (Lander et al. 2001; Waterston et al. 2002) and were initially classified as “junk” DNA (Ohno 1972). While transposable elements are still regarded as the main danger for genome integrity (Doolittle & Sapienza 1980; Orgel & Crick 1980), their function as drivers of genome evolution and fundamental regulatory roles has recently begun to be addressed. Transposable elements can be classified in two major classes according to their mechanism of transposition: class I elements or retrotransposons, which transpose through an RNA intermediate, and class II elements or DNA transposons that can directly move from one site in the genome to another as DNA molecules (reviewed in García-Pérez et al. 2016) (Figure 1.3). Retrotransposons can be further classified into endogenous retroviruses (ERVs), long interspersed elements (LINEs) and short interspersed elements (SINEs) (Okamura & Nakai 2008) (Figure 1.3). Approximately 3% of mammalian genomes are made of DNA transposons, which are no longer mobile in most species (Lander et al. 2001; Waterston et al. 2002). On the contrary, retrotransposons can constitute up to 40% of a typical mammalian genome and are generally active (Lander et al. 2001); (reviewed in reviewed in Richardson et al. 2015). During mouse and human pre-implantation development, retrotransposons are expressed and contribute to cellular identity

(X. Lu et al. 2014; Jachowicz et al. 2017) and ZGA (Jachowicz et al. 2017; De Iaco et al. 2017; Hendrickson et al. 2017; Whiddon et al. 2017) (reviewed in Rowe & Trono 2011; Rodriguez-Terrones & Torres-Padilla 2018).

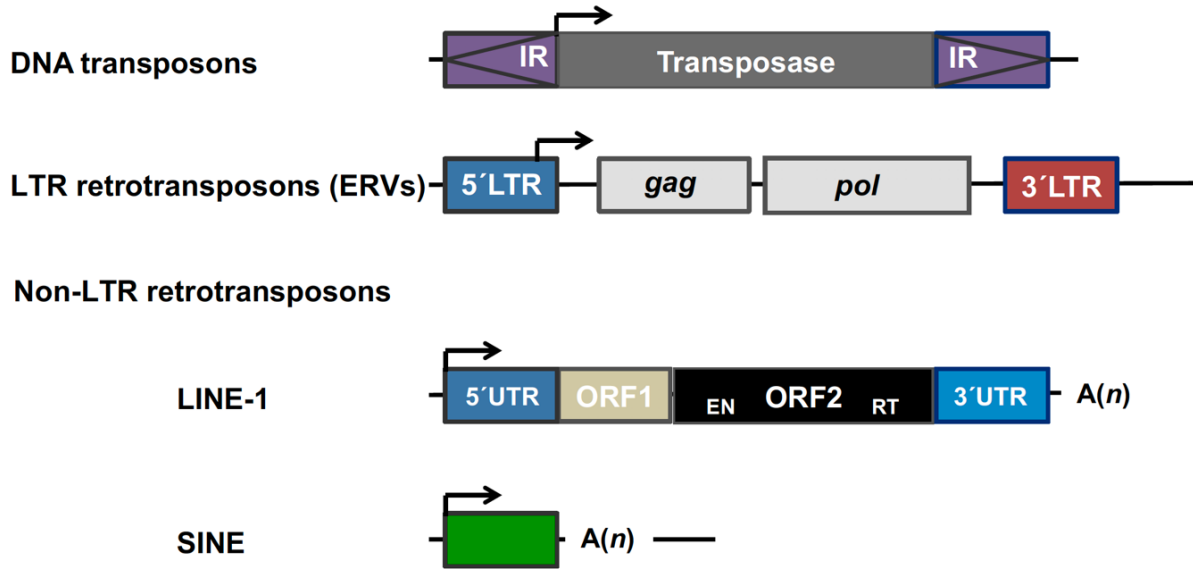



**Figure 1.3– Structures of key classes of transposable elements in the mammalian genome**

DNA transposons encode a transposase gene flanked by tandem inverted repeats (IRs). LTR retrotransposons or endogenous retroviruses (ERVs) encode the genes *Gag* and *Pol* flanked by long terminal repeats (LTRs). Two key families of non-LTR retrotransposons are LINE-1 and SINE elements. LINE-1 elements contain an internal promoter at their 5' end, a variable number of open reading frames (ORFs) and a 3' untranslated region (3' UTR) with a polyadenylation signal and a polyA tail. The structure of SINE elements varies for different sub-families, but their 5' ends are generally derived from small cellular RNAs and they encode a polyA tail. Figure adapted from García-Pérez et al. 2016.

ERVs contain long terminal repeat (LTR) sequences equally oriented at both ends of the retrotransposon, which provide a promoter at the 5' end and a transcriptional termination and polyadenylation signals at the 3' end (Mager & Stoye 2015) (Figure 1.3). A full-length ERV typically encodes two genes: *Gag*, a polymorphic capsid protein with structural and nucleic acid-binding functions, and *Pol*, which encodes a protease, a reverse transcriptase, an RNAseH domain and an integrase (Wicker et al. 2007) (Figure 1.3). ERVs can be autonomous, that is, able to encode functional retroviral proteins required for mobilisation, or non-autonomous, which rely on retroviral proteins encoded by active ERVs to mobilize (reviewed in García-Pérez et al. 2016). ERVs are expressed during mouse MZT. For example, mammalian apparent

LTRs (MaLRs) are contained in approximately 12% of mouse oocyte transcripts (Peaston et al. 2004). Murine ERVs with leucine tRNA primer-binding site (MERVL) elements are expressed in one- and two-cell embryos (Kigami 2002; Peaston et al. 2004; Svoboda et al. 2004; Macfarlan et al. 2012). MERVL elements can be full-length or composed of a *solo* LTR (also known as MT2 elements) (Wicker et al. 2007). There are 656 MERVL full-length copies and 37,172 MT2 copies in the C57Bl/6 genome and its transcripts account for nearly 4% of the mouse transcriptome at the two-cell stage (Kigami 2002; Peaston et al. 2004; Svoboda et al. 2004). MERVL elements often function as alternative promoters during ZGA, enabling transcription of neighboring genes that results in chimeric transcripts and allows coordinated expression during the major wave of activation (Peaston et al. 2004; Y. Huang et al. 2017; Macfarlan et al. 2012; Franke et al. 2017). Additionally, MERVL elements can derive non-coding RNAs (Franke et al. 2017; Karlic et al. 2017) and drive evolution of new protein-coding genes during ZGA (Karlic et al. 2017). The human counterpart, HERVL elements, also drive ZGA at the eight-cell stage embryo (De Iaco et al. 2017; Hendrickson et al. 2017; Pontis et al. 2019). Furthermore, through evolution, these elements can form novel binding sites for transcription factors, both in mouse and human, providing binding sites for regulators of ZGA (Bourque et al. 2008; Kunarso et al. 2010; Imbeault et al. 2017; De Iaco et al. 2017; Hendrickson et al. 2017; Whiddon et al. 2017; Pontis et al. 2019).

LINE elements are autonomous transposable elements that lack LTRs and comprise approximately between 10% to 30% of mammalian genomes. L1 elements are the most abundant LINE elements in mouse and human (reviewed in Penzkofer et al. 2017). The full-length elements contain an internal promoter at their 5' end, a variable number of open reading frames (ORFs) and a 3' untranslated region (3' UTR) with a polyadenylation signal and a polyA tail (Figure 1.3). Young L1 elements are transcribed in one-cell mouse embryos and regulate chromatin accessibility during ZGA (Beraldi et al. 2006; Fadloun et al. 2013; Jachowicz et al. 2017; Percharde et al. 2018), specifically, they repress the zygotic transcription factor DUX, allowing exit from the two-cell stage (Percharde et al. 2018) (see sections 1.1.4.2 and 1.1.6.4-1.1.6.5).

SINE elements are non-autonomous transposable elements that rely on the enzymatic machinery of LINE elements for their transposition (reviewed in Richardson et al. 2015). Most

SINE elements have a 5' end derived from small cellular RNAs transcribed by polymerase III, a central sequence of diverse origin and a 3' tail related to the 3' end of a LINE (Figure 1.3). While their functional role, if any, during ZGA is unknown, the B2 subfamily of SINE elements is transcribed in mouse zygotes during the minor wave of ZGA (Vasseur et al. 1985).

Besides transposable elements, there are other repetitive elements in the genome whose transcription is active during ZGA. This is the case for major satellites, pericentric heterochromatin repeats with AT-rich long repetitive domains (Wong & Rattner 1988; Joseph et al. 1989). Major satellites are actively transcribed in two-cell embryos, where they play a role in pericentromeric heterochromatin organisation (Probst et al. 2010; Casanova et al. 2013) (see section 1.1.4.2).

### 1.1.4 Epigenetic dynamics during the maternal-to zygotic transition

The term "epigenetics" has been given multiple definitions since it was first coined by Waddington in 1942. Here, I use "epigenetics" to refer to "the layer of information that exists beyond that encoded in the DNA sequence itself, thereby making the genome function distinctively in different cell types" (reviewed in Greally 2018). Epigenetic regulation is critical for normal development and its deregulation can be associated with various diseases (reviewed in Portela & Esteller 2010; Reik et al. 2003). The epigenome is extensively remodeled during the MZT, by changes in global DNA methylation, histone post-translational modifications (PTMs), local chromatin remodeling and higher-order chromatin architecture (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018; Jansz & Torres-Padilla 2019; Ladstatter & Tachibana 2019). Epigenetic remodeling, together with the associated changes in the gene expression landscape, changes cellular identity and confers totipotency to the embryo (reviewed in Ishiuchi & Torres-Padilla 2013; Ladstatter & Tachibana 2019).

#### 1.1.4.1 Global DNA demethylation

DNA methylation is a covalent modification of the DNA itself by which a methyl group is added to the 5-carbon of the cytosine base (5mC) (Holliday & Pugh 1975; Razin & Riggs 1980) (Figure 1.4). In the mammalian genome, 5mC is mostly found in cytosine residues of CpG

dinucleotides (Bird et al. 1985), although non-CpG methylation is also present at low levels, mostly in the CpA context (Bird 2002). Approximately 60% to 80% of all CpG sites in the mammalian genome are methylated, with the notable exception of regions rich in the CpG dinucleotide, known as CpG islands, which are mostly unmethylated and found in promoters of active genes, particularly housekeeping and developmental genes (Bird 2002). CpG methylation at gene promoters has long been associated to transcriptional repression (Razin & Riggs 1980) via different mechanisms such as recruitment of histone post-translational modifications or by preventing the binding of transcription factors (reviewed in Klose & Bird 2006). DNA methylation is established and maintained by DNA methyltransferases (DNMTs). Specifically, the maintenance DNMT1, together with its interacting partner E3 ubiquitin-protein ligase UHRF1, methylate the unmodified nascent DNA strand opposite to a previously methylated position after DNA replication (reviewed in Bestor 2000; E Li et al. 1992) (Figure 1.4). The *de novo* methyltransferases DNMT3A and DNMT3B, regulated by the catalytically inactive isoform DNMT3L, establish new DNA methylation patterns (reviewed in Bestor 2000); (Okano et al. 1999; Barau et al. 2016) (Figure 1.4). Removal of DNA methylation can occur either passively by dilution through impaired maintenance upon DNA replication, or actively via DNA repair pathways and/or oxidation reactions catalysed by the ten-eleven translocation (TET) enzymes (reviewed in H. Wu & Yi Zhang 2014) (Figure 1.4).

During mammalian development, the genome undergoes two waves of global DNA demethylation: during the establishment of PGCs and in the zygote after fertilisation (reviewed in H. J. Lee et al. 2014) (Figure 1.5). Regulation of DNA methylation in these two critical processes plays a major role in genome regulation, genome stability, X chromosome inactivation, transposon silencing and genomic imprinting (Reik & Walter 2001; reviewed in Greenberg & Bourc'his 2019). Genomic imprinting refers to the epigenetic phenomenon driven by DNA methylation in which certain genes are solely expressed from either the paternal or maternal allele and retain their DNA methylation status (parental imprints) during epigenetic reprogramming in the early embryo (Figure 1.5). For the purpose of this dissertation, I next review the DNA methylation dynamics that occurs after fertilisation and through pre-implantation development.

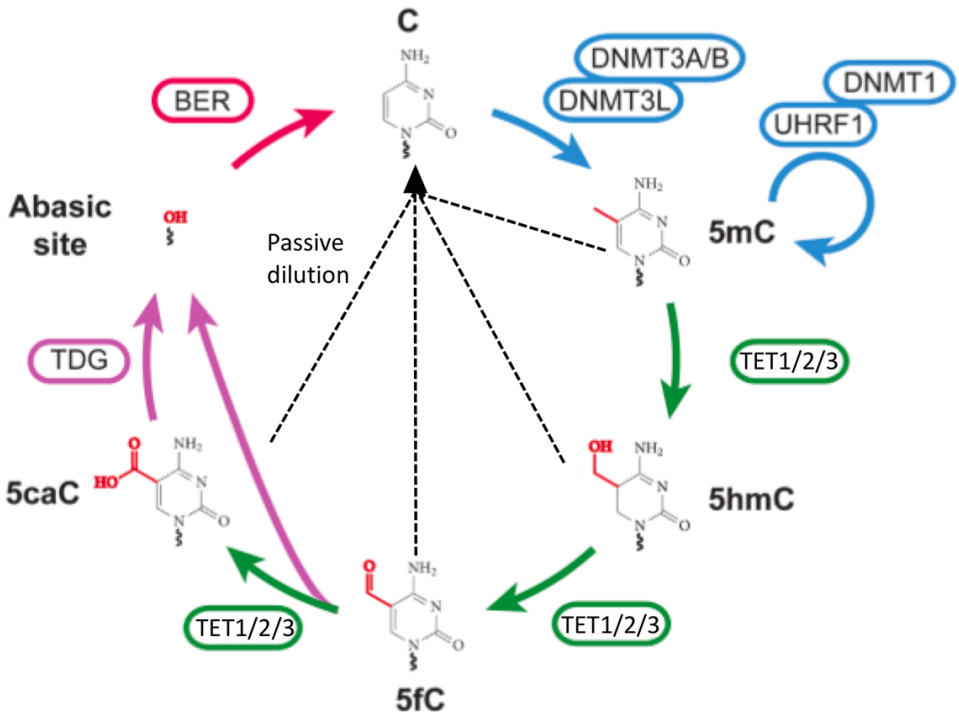



**Figure 1.4.– DNA methylation and demethylation machinery**

Unmodified cytosines are *de novo* methylated by the DNA methyltransferase enzymes DNMT3A/B regulated by the co-factor DNMT3L. DNMT1, together with UHRF1, maintains DNA methylation patterns across DNA replication. TET enzymes (TET1/2/3) oxidise 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5' formylcytosine (5fC) and 5-carboxycytosine (5caC). 5fC and 5caC can be excised by thymine-DNA glycosylase (TDG), leaving an abasic site that triggers base-excision repair (BER) mechanisms. DNA demethylation can also occur through passive dilution of 5mC, 5hmC, 5fC or 5caC. Figure modified from H. Wu & Yi Zhang 2014.

DNA methylation reprogramming in the zygote erases the epigenetic "memory" inherited from the parents, with the notable exception of parental imprints, and allows re-establishment of developmental totipotency (reviewed in H. J. Lee et al. 2014). This process occurs differentially in the maternal and paternal genome (Figure 1.5), as demonstrated both by immunofluorescence-based and genome-wide analysis (Oswald et al. 2000; Mayer et al. 2000; Wossidlo et al. 2010; Wossidlo et al. 2011; Smith et al. 2012; L. Wang et al. 2014; F. Guo et al. 2014; Smith et al. 2014; Peat et al. 2014; Santos et al. 2013; Amouroux et al. 2016; Petrussa et al. 2016).

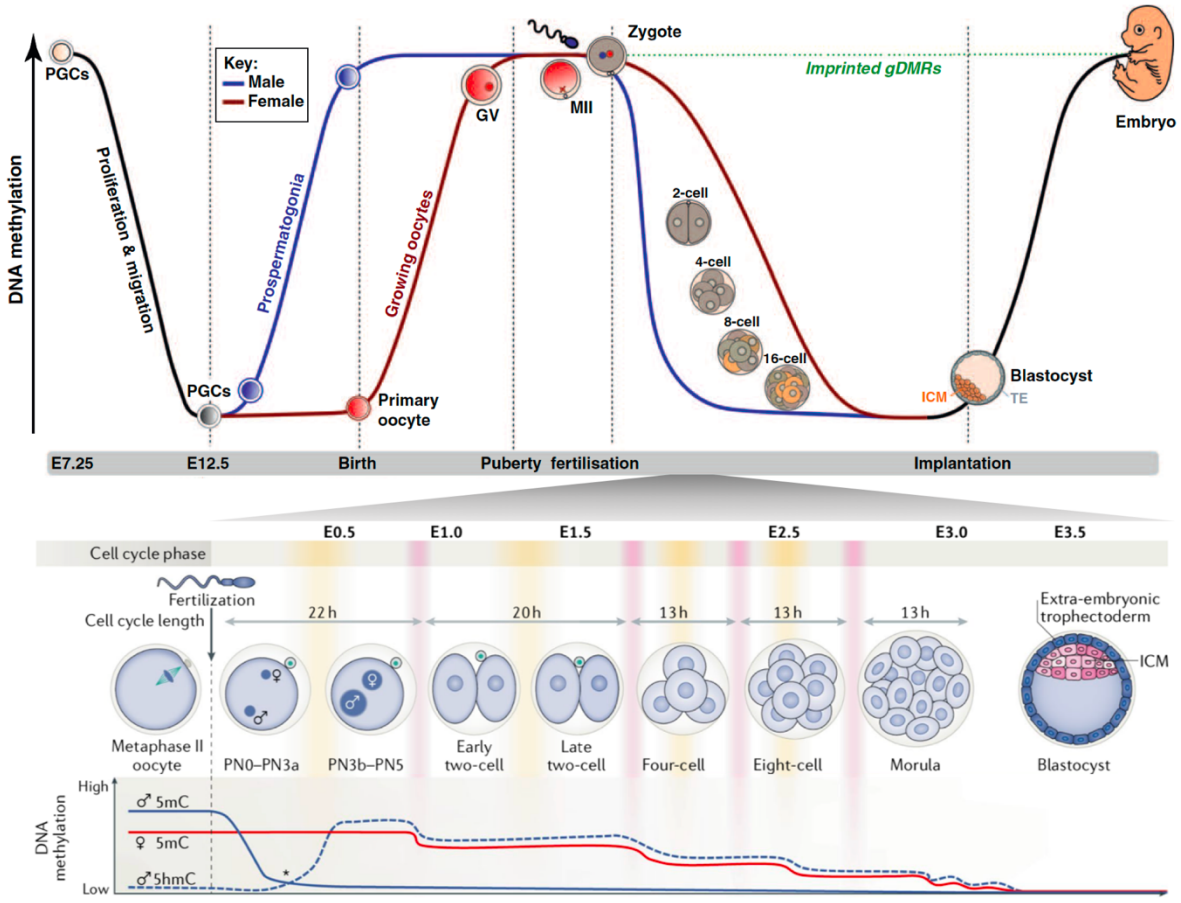



**Figure 1.5– Overview of CpG methylation dynamics during mouse development**

In primordial germ cells (PGCs), reprogramming is necessary to erase the DNA methylation pattern originated in the epiblast, from which they are derived at E7.25, to reset parental imprints in a gender-specific manner, and to gain developmental potency. Concomitant with proliferation and migration towards the genital ridge, PGCs undergo a wave of DNA demethylation from E7.25 to E12.5 (black line, top panel). After sex determination, new DNA methylation patterns are established in the germ cell precursors in a sex-specific manner (blue and red lines, top panel). Following fertilisation, a new wave of DNA demethylation takes place differentially in the paternal and maternal genome throughout pre-implantation development. Paternal DNA methylation (5mC, blue solid line, top and bottom panel) is rapidly lost by mechanisms of active demethylation. Paternal 5-hydroxymethylation (5hmC, blue dashed line, bottom panel) increases rapidly after fertilisation, while *de novo* DNA methylation also occurs (*, bottom panel). Maternal DNA methylation (red solid line, top and bottom panel) is lost by passive dilution following DNA replication during cleavage divisions. This wave of DNA demethylation after fertilisation does not affect parental imprints (imprinted genomic differentially methylated regions, gDMRs, green dashed line, top panel). After implantation of the blastocyst into the mother's uterus, DNA methylation patterns are re-established associated with lineage determination. In the bottom panel, approximate cell cycle phases are represented by yellow bars for S phase and red bars for M phase. Approximate embryonic days (E) are denoted. Image adapted from Smallwood & Kelsey 2012 (top panel) and Eckersley-Maslin, Alda-Catalinas & Reik 2018 (bottom panel).

CpG methylation is established in mouse oocytes by the DNMT3A/DNMT3L complex in around 40% of CpG sites (Smallwood et al. 2011; Kobayashi et al. 2012). Interestingly, and unlike the genome of sperm and somatic cells, the oocyte is hypomethylated in transcriptionally inert regions (Smallwood et al. 2011; Kobayashi et al. 2012) (reviewed in Stewart et al. 2016), a feature that is maintained, at least in part, by STELLA (also known as DPPA3 or PGC7) and is essential for development after fertilisation and ZGA (Y. Li et al. 2018). DNA methylation in the maternal genome is passively lost in a replication-dependent manner upon fertilisation and during pre-implantation development (Rougier et al. 1998) (Figure 1.5), predominantly by nuclear exclusion of DNMT1 in the female pronucleus (Howell et al. 2001; Mahadevan et al. 2017) and cytoplasmic localisation of UHRF1 (Maenohara et al. 2017). However, low levels of active maternal DNA demethylation were also recently observed in zygotes (F. Guo et al. 2014; Shen et al. 2014).

Mouse sperm presents high levels of DNA methylation, around 90% of CpGs, also established by the DNMT3A/DNMT3L complex (Erkek et al. 2013). After fertilisation, the male pronucleus rapidly demethylates by mechanisms of active demethylation (Figure 1.5), through oxidation of 5mC to 5-hydroxymethylation (5hmC), 5-formylmethylation (5fC) and 5-carboxymethylation (5caC), as demonstrated by immunofluorescence (T.-P. Gu et al. 2011; Inoue & Yi Zhang 2011; Iqbal et al. 2011; Santos et al. 2013; Wossidlo et al. 2011). This oxidation is partly driven by TET3, the only TET enzyme expressed one-cell embryos and whose knock-down or maternal knock-out (KO) inhibits DNA demethylation in the male pronucleus (T.-P. Gu et al. 2011; Wossidlo et al. 2011; Inoue et al. 2012; Peat et al. 2014; F. Guo et al. 2014; Inoue et al. 2015; Tsukada et al. 2015). Moreover, TET3-mediated demethylation is also responsible for the transcriptional activation of transposable elements (Inoue et al. 2012). However, recent studies have shown that 5hmC and 5fC only rise in the late pronuclear stages of the male genome (Santos et al. 2013; Amouroux et al. 2016; Zhu et al. 2017) (Figure 1.5), suggesting that active demethylation involves other mechanisms other than TET3 which are currently unknown. Consequently, it has been suggested that TET3 also has a role in counteracting the *de novo* DNA methylation observed in zygotes (Amouroux et al. 2016) (Figure 1.5). Other machinery that might contribute to the demethylation of the male pronucleus is the maternal factor gonad-specific expression gene (GSE), whose depletion

reduces the loss of 5mC and gain of 5hmC in the paternal genome (Hatanaka et al. 2013). Additionally, base-excision repair (BER) mechanisms contribute to paternal DNA demethylation by replacing modified cytosines with unmodified ones (Wossidlo et al. 2010; Hajkova et al. 2010; Wossidlo et al. 2011; Santos et al. 2013; Ladstatter & Tachibana-Konwalski 2016) (Figure 1.4). How these different mechanisms act together to rapidly demethylate the paternal pronucleus is an active area of research. For example, it was shown that TET3-mediated demethylation generates DNA lesions in the paternal pronucleus that are repaired by a cohesin-dependent mechanism, preventing activation of a mitotic checkpoint and allowing cleavage divisions to proceed (Ladstatter & Tachibana-Konwalski 2016).

The maternal genome and paternal imprints are protected from TET3-mediated demethylation by preferential binding of STELLA to dimethylated histone H3 at Lysine 9 (H3K9me2) (Hajkova et al. 2010; Nakamura et al. 2012), which is not found in the paternal pronucleus (Santos et al. 2005). Other recently proposed candidates to establish the asymmetry in active DNA demethylation between the paternal and maternal pronuclei are the H3K9 methyltransferases euchromatic histone lysine methyltransferase 2 (EHMT2, also known as G9a) and SET domain bifurcated histone lysine methyltransferase 1 (SETDB1) (T.-B. Zeng et al. 2019). Imprinted regions are also maintained by the cooperative action of zinc-finger protein 57 (ZFP57) and KRAB-associated protein-1 (KAP1), which are able to recruit DNMT1 to these loci, ultimately driving maintenance of methylation (Xiajun Li et al. 2008; Quenneville et al. 2011).

In human embryos, the major wave of DNA demethylation is completed by the two-cell stage, which is surprisingly earlier than what is observed in mouse embryos. However, and consistent with the mouse model, the paternal genome demethylates much faster than the maternal pronuclei (H. Guo et al. 2014). Nevertheless, the mechanistic insights of how differential DNA methylation is regulated between the two parental genomes at this stage remains unknown.

### 1.1.4.2 Chromatin remodeling

Chromatin refers to the molecular organisation of DNA and proteins in the nucleus of a cell (Flemming, 1882). The basic structural unit of the eukaryotic chromatin is the nucleosome

(Figure 1.6). Each nucleosome contains eight histone proteins (2 copies of H2A, H2B, H3 and H4) and approximately 147 base pairs (bp) of DNA, which, via another histone protein, H1, can be packed into different levels of structural organisation (reviewed in D. E. Olins & A. L. Olins 2003). Nucleosome composition and histone post-translational modifications (PTMs) determine the degree of chromatin compaction, which subsequently influences transcriptional activity through alterations in transcriptional factor binding and affinity of enzymatic machinery (reviewed in Venkatesh & Workman 2015; Allis & Jenuwein 2016; Lawrence et al. 2016; Harr et al. 2016) (Figure 1.6). According to the level of compaction, mammalian genomes are largely organised in two distinct chromatin states: euchromatin, a state generally permissive for gene activation, and heterochromatin, which is largely gene poor and transcriptionally silenced (reviewed in Grewal & Elgin 2002). During mammalian MZT, the chromatin landscape is highly remodelled via changes in histone PTMs, local chromatin accessibility and higher-order genome organisation (Figure 1.7), which together with the reprogramming of DNA methylation after fertilisation, set a new epigenetic landscape in the early embryo (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018; Jansz & Torres-Padilla 2019; Ladstatter & Tachibana 2019).

#### 1.1.4.2.1 Global dynamics of histone post-translational modifications

In mammals, inactive genes are generally marked by methylation of histone H3 lysine 9 (H3K9), histone H3 lysine 27 (H3K27) and histone H4 lysine 20 (H4K20), albeit active genes are marked with high levels of acetylation and methylation of histone H3 lysine 4 (H3K4), histone H3 lysine 36 (H3K36) and histone H3 lysine 79 (H3K79) (reviewed in Kouzarides 2007) (Figure 1.6). Trimethylation of H3K4 (H3K4me3) is usually found in promoter regions whereas trimethylation of H3K36 (H3K36me3) is found in gene bodies of actively transcribed genes (Liang et al. 2004). Many genes important for development are termed “poised” because they present co-occurrence of the activating H3K4me3 and the repressing trimethylation of H3K27 (H3K27me3) marks at their promoters, a feature that is called bivalent chromatin and allows rapid gene expression upon lineage specification (Bernstein et al. 2006; Azuara et al. 2006). Additionally, monomethylation of H3K4 (H3K4me1) and acetylation of H3K27 (H3K27ac) mark cell type specific enhancers (Rada-Iglesias et al. 2011), while a subset of

enhancers has also been found to be marked by acetylation of histone H3 lysine 122 (H3K122 ac) and lack of H3K27ac (Pradeepa et al. 2016).

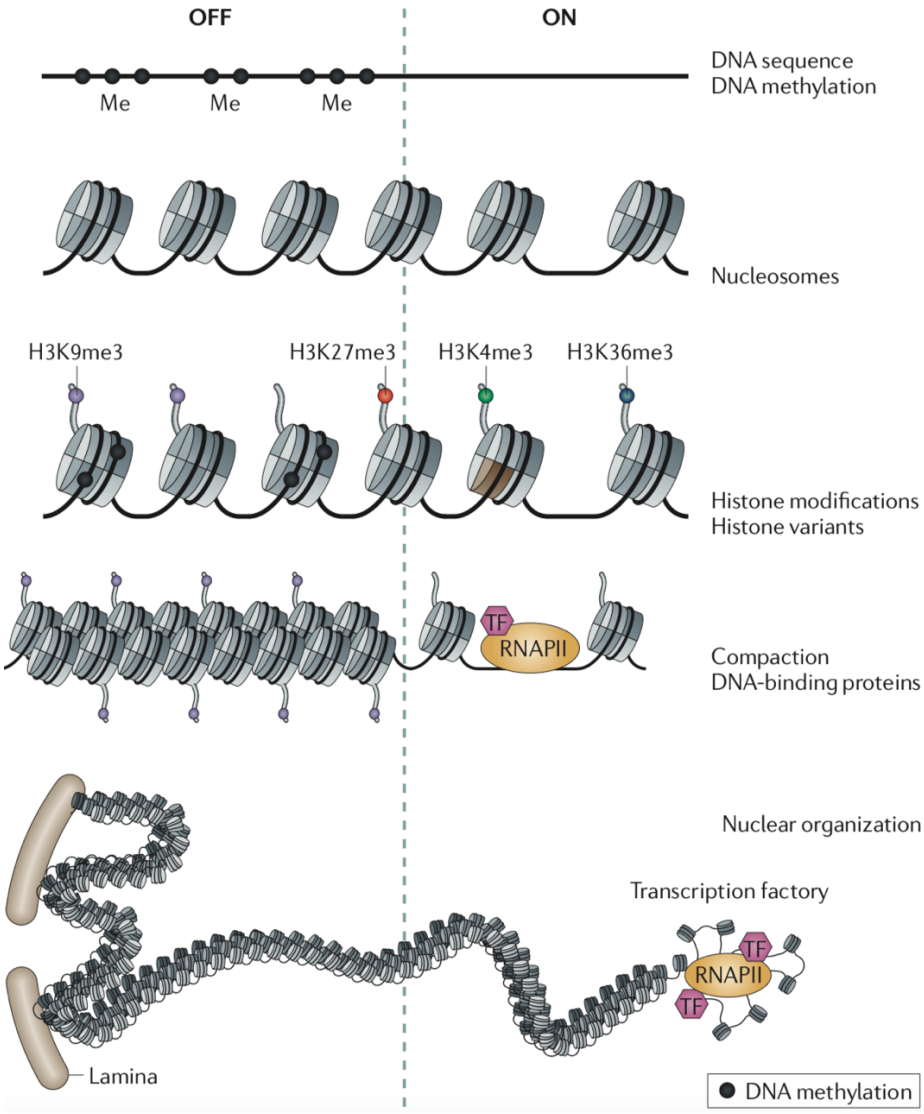



**Figure 1.6– Chromatin organisation in the mammalian genome**

In mammals, DNA is associated with histone proteins to form chromatin. The basic structural units of chromatin are nucleosomes, which can be found into more or less compacted chromatin based on DNA and histone post-translational modifications. Chromatin can also organise into higher-order structures such as lamina-associated domains (LADs). The transcriptional state of chromatin can be inactive (off) or active (on) as determined by DNA methylation and histone post-translational modifications, which in turn determine accessibility to DNA binding proteins such as transcription factors (TF) and RNA polymerase II (RNAPII). DNA methylation, trimethylation of histone H3 lysine 9 (H3K9me3) and trimethylation of histone H3 lysine 27 (H3K27me3) are generally associated with transcriptional repression, whereas trimethylation of histone H3 lysine 4 (H3K4me3) and trimethylation of histone H3 lysine 36 (H3K36me3) are generally associated with active transcription. Image adapted from Zhou, Goren & Bernstein 2011.

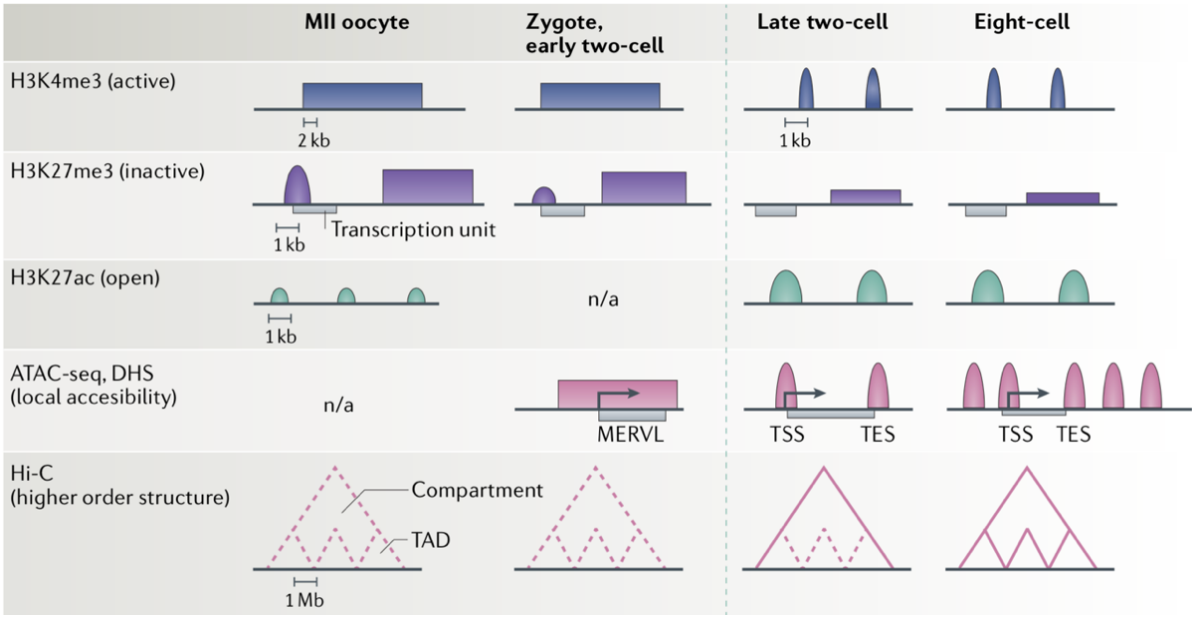



**Figure 1.7– Chromatin remodeling during the MZT**

The pattern of different chromatin features from oocytes to two-cell embryos is shown. The chromatin marks H3K4me3 (blue), H3K27me3 (purple) and H3K27ac (green) have been profiled by Chromatin Immunoprecipitation-sequencing (ChIP-seq). H3K4me3 (blue), a mark associated with transcriptional activity, is present in MII oocytes as broad domains, non-overlapping with similarly broad domains found for the H3K27me3 mark (purple), associated with transcriptional repression. These broad peaks of both H3K4me3 and H3K27me3 are maintained in zygotes and early two-cell embryos but disappear after the major wave of ZGA (dashed vertical line). After ZGA, canonical peaks of H3K4me3 are established. H3K27me3 is also present in a canonical peak structure in oocytes, but these disappear after ZGA. H3K27ac (green), associated with open chromatin, is gradually gained after ZGA. The local chromatin structure (pink) has been measured by Assay for Transposable Accessible Chromatin followed by sequencing (ATAC-seq) and by DNaseI hypersensitivity sequencing (DHS) in several stages after fertilisation, assays that revealed large domains of accessible chromatin covering MERVL elements before the major wave of ZGA. From the late two-cell embryo, after ZGA, the number of accessible regions increases both at transcription start sites (TSS) and at transcription end sites (TES). Higher order chromatin structure has been measured by chromatin conformation capture-based (Hi-C) approaches showing compartments and topologically associating domains (TADs) disordered in MII oocytes, zygotes and early two-cell embryos. After ZGA, compartments are established and subsequently, TADs. Pink dashed lines represent disordered or unstructured features and solid pink lines represent established features. n/a: data not available. Image from Eckersley-Maslin, Alda-Catalinas & Reik 2018.

While histone PTMs in early mammalian development have traditionally been studied by immunofluorescence (reviewed in Burton & Torres-Padilla 2010), the recent development of low-input chromatin immunoprecipitation followed by sequencing (ChIP-seq) techniques has allowed genome-wide profiling of H3K4me3, H3K27me3 and H3K27ac during mouse MZT

and pre-implantation development (Dahl et al. 2016; Xiaoyu Liu et al. 2016; B. Zhang et al. 2016; Hanna et al. 2018) (Figure 1.7). Additionally, using CUT&RUN technique (Skene & Henikoff 2017), H3K4me3 and H3K27me3 has also been profiled in human embryos (Xia et al. 2019)

In mouse MII oocytes, H3K4me3 is present in large blocks spanning more than 10 kilobases (kb) that cover approximately 22% of the genome (Figure 1.7), a unique pattern very distinct from the sharp peaks typically present at transcription start sites (TSS) in somatic cells (Dahl et al. 2016; Xiaoyu Liu et al. 2016; B. Zhang et al. 2016; Hanna et al. 2018). These blocks of H3K4me3 are gradually established during oogenesis and anti-correlate with DNA methylation (Smallwood et al. 2011; Dahl et al. 2016; B. Zhang et al. 2016; Hanna et al. 2018), in fact, depletion of the *de novo* DNA methyltransferases DNMT3A and DNMT3B showed that DNA methylation protects regions from acquiring H3K4me3 blocks (Hanna et al. 2018). Interestingly, even though H3K4me3 domains are not broadly associated with gene promoters in oocytes, the promoters of genes expressed during the major wave of ZGA are pre-marked by H3K4me3 both in oocytes and in sperm (Dahl et al. 2016). Upon fertilisation, these broad domains are maintained in the maternal, but not in the paternal, pronucleus of the zygote and early two-cell embryos, but from the late two-cell embryo, they are depleted and classical sharp peaks are established (Dahl et al. 2016; Xiaoyu Liu et al. 2016; B. Zhang et al. 2016; Lepikhov & Walter 2004) (Figure 1.7). These parental-specific differences likely reflect the genome-wide replacement of sperm-derived protamines with histones upon fertilisation (Nonchev & Tsanev 1990). From the late two-cell stage, promoter H3K4me3 correlates with transcription, suggesting that H3K4me3 specifies permissive promoters of either recently transcribed genes or genes poised for activation (B. Zhang et al. 2016). Concomitant with the loss of H3K4me3 levels from oocytes to two-cell embryos, H3K27ac increases markedly from the oocyte to the eight-cell embryo (Figure 1.7), especially in major ZGA gene TSS-containing domains previously marked by the broad H3K4me3 in oocytes (Dahl et al. 2016). Strikingly, in human GV and metaphase I (MI) oocytes, H3K4me3 is present in canonical promoter peak patterns rather than in broad domains, a landscape that is maintained until the four-cell stage before ZGA (Xia et al. 2019). Similar to mouse embryos, approximately half of the promoters marked

with H3K4me3 in pre-ZGA stages retain this mark and become activated during ZGA in eight-cell stage embryos (Xia et al. 2019).

H3K27me3 is also present as broad domains in mouse MII oocytes (Figure 1.7) in regions depleted of transcription, H3K4me3 and DNA methylation (H. Zheng et al. 2016). These broad distal domains of H3K27me3 are present in two-cell embryos during ZGA and also in the ICM of E3.5 blastocysts (Figure 1.7) and persist until the post-implantation embryo, where they are converted to canonical H3K27me3 peaks marking bivalent promoters (Xiaoyu Liu et al. 2016; H. Zheng et al. 2016). Overall, bivalent promoters are infrequent and unstable in early embryos compared to embryonic stem cells (ESCs) and post-implantation embryos (Xiaoyu Liu et al. 2016). However, H3K27me3 is also found in oocytes with a canonical peak pattern in promoters of Hox genes and other developmentally-relevant genes (H. Zheng et al. 2016) (Figure 1.7). There are also promoter-distal H3K27me3 peaks in oocytes that are inherited by the zygote (Figure 1.7) whereas promoter-associated peaks at developmental genes disappear before major ZGA (H. Zheng et al. 2016). In contrast, most H3K27me3 sperm-derived peaks are lost in zygotes following fertilisation (H. Zheng et al. 2016), consistent with immunostainings showing H3K27me3 detection in the maternal but not in the paternal pronucleus (Burton & Torres-Padilla 2010). *De novo* paternal-specific H3K27me3 peaks are then established in late pre-implantation stages and throughout post-implantation in genes that remain silent (H. Zheng et al. 2016). Similarly, in human oocytes, H3K27me3 is present both in canonical peak promoters of developmental genes and also in partially-methylated domains, but unlike mouse maternal H3K27me3, human embryos at the time of ZGA are nearly depleted of this histone mark (Xia et al. 2019).

As explained above, DNA demethylation mechanisms are drastically different in the maternal and paternal genome (see section 1.1.4.1). This might be explained by the different chromatin environments of the two pronuclei given by the differences observed in histone variants and histone PTMs, including lysine methylation marks in histones H3 and H4 and histone H4 acetylation (reviewed in Burton & Torres-Padilla 2010; Zhou & Dean 2015).

#### 1.1.4.2.2 Local chromatin dynamics

The chromatin accessibility landscape has also been profiled from mouse zygotes throughout pre-implantation using both Assay for Transposable Accessible Chromatin followed by sequencing (ATAC-seq) (J. Wu et al. 2016) and DNaseI hypersensitivity (DHS) mapping (F. Lu et al. 2016; Inoue et al. 2017) (Figure 1.7). ATAC-seq relies on the use of a hyperactive Tn5 transposase to label open chromatin regions (Buenrostro et al. 2013), whereas DHS mapping or DNase-seq identifies open chromatin sites as those sensitive to DNaseI cleavage (Crawford et al. 2006). In zygotes, while DHS mapping showed peaks mostly located at gene promoters (F. Lu et al. 2016; Inoue et al. 2017), ATAC-seq revealed a weak and noisy profile where large domains of open chromatin were covering genes expressed during the minor wave of ZGA, including MERVL elements (J. Wu et al. 2016) (Figure 1.7), consistent with the pervasive transcription observed at this stage (see section 1.1.3.1). Following major ZGA, at the late two-cell stage, ATAC-seq showed peaks of accessible chromatin at TSSs and transcription end sites (TES) (J. Wu et al. 2016), the functional relevance of which is unclear. After the MZT is completed, from the eight-cell stage onwards, the overall landscape of chromatin becomes more accessible (J. Wu et al. 2016), with a drastic increase of DHS peaks gained at the eight-cell stage (Figure 1.7) at promoters, intragenic and intergenetic regions (F. Lu et al. 2016). This increase in chromatin accessibility correlates with an increase in transcriptional activity (F. Lu et al. 2016), however, accessible promoters are also found in inactive genes that become activated at later developmental stages, suggesting these promoters are primed for activation (F. Lu et al. 2016). Similarly, in human embryos, promoter accessibility correlates with major ZGA and, overall, there is a global increase in accessible regions during cleavage stages (L. Gao et al. 2018; Lin Li et al. 2018).

Marked differences have been found in chromatin accessibility between the paternal and the maternal genome. After fertilisation, the paternal genome experiences the most extensive reprogramming since the highly packaged chromatin in protamines of the sperm needs to be replaced with maternal histones (Nonchev & Tsanev 1990; van der Heijden et al. 2005; Torres-Padilla et al. 2006). Despite this exchange and the drastic differences in the DNA methylation and histone PTMs landscapes between the two pronuclei after fertilisation (see sections 1.1.4.1 and 1.1.4.2.1), the local chromatin structure, as analysed by DHS mapping, does not present

major differences between the paternal and maternal genome (F. Lu et al. 2016). Interestingly though, DHS parental-specific peaks have been found to be priming genes in the zygote that are later on allelically expressed at the time of major ZGA (Inoue et al. 2017). Furthermore, paternal-specific DHS-marked genes have been found to be marked with H3K27me3 in the corresponding maternal allele in zygotes (Inoue et al. 2017). Later on, after the two-cell stage, the DHS profiles are similar between parental alleles, except imprinted genes that present a differential chromatin accessibility profile before becoming expressed (F. Lu et al. 2016). On the contrary, in human embryos, the chromatin accessibility, as measured by the single-cell multi-omics approach "chromatin overall omic-scale landscape sequencing" (scCOOL-seq) is higher in the paternal than in the maternal genome from mid-zygotic stages to the four-cell stage (Lin Li et al. 2018), likely reflecting a paternal-specific minor wave of ZGA.

#### 1.1.4.2.3 High-order chromatin architecture

Traditionally, nuclear sub-structures, such as the nucleolus, nuclear envelope and lamina, promyelocytic leukemia and Cajal bodies, nuclear speckles, euchromatin and heterochromatin, and chromosome territories have been studied with microscopy-based techniques in pre-implantation mouse embryos (reviewed in Borsos & Torres-Padilla 2016). A marked difference in early embryos compared to somatic cells is the organisation of major satellite repeats, which comprise pericentromeric heterochromatin, around nucleolar precursor bodies (NPBs) (Probst et al. 2007), whose localisation is critical for the establishment of repressive heterochromatin and further development (Jachowicz et al. 2013). Of special interest as well is the recent development of high-resolution maps of genomic interactions with the nuclear lamina (Borsos et al. 2019), a fibrillar network in the inner nuclear membrane (Figure 1.6), using the DNA adenine methyltransferase identification (DamID) technique (van Steensel et al. 2001). These maps revealed that lamina-associated domains (LADs) are not inherited from the oocyte but rather established *de novo* after fertilisation in a parental-specific and DNA replication-independent manner, with fewer genome-nuclear lamina contacts in the maternal than in the paternal genome at the zygote stage (Borsos et al. 2019). Interestingly, ectopic expression of the lysine-specific demethylase 5B (KDM5B, also known as JARID1B or PLU-1), which demethylates H3K4, inhibited paternal LAD formation in zygotes, suggesting that LAD

establishment may be dependent on H3K4me3 remodeling. LADs AT content changes substantially from zygotes to two-cell embryos (Borsos et al. 2019), possibly as a result of major genome reorganisation at the time of ZGA. At this stage, maternal genome-nuclear lamina contacts are reinforced whereas paternal genome-nuclear lamina contacts are rearranged (Borsos et al. 2019). After the eight-cell stage, parental-specific differences disappear and LADs become more strongly associated with reduced chromatin accessibility sites (Borsos et al. 2019).

Recent advances in low-input chromatin conformation captured-based (Hi-C) techniques have allowed the development of multi-hierarchical chromatin structure maps in early mouse embryos, describing the remodeling and establishment of loops, topologically associating domains (TADs) and compartments during the MZT (Du et al. 2017; Flyamer et al. 2017; Gassler et al. 2017; Ke et al. 2017). Loops are three-dimensional (3D) structures formed by close contact of two distal genomic regions. TADs refer to regions of approximately 400-500 kilobase (kb) in length that present high contact frequency and contain loci that are more likely to interact with each other than across TAD boundaries. TADs further organize into spatially segregated regions of approximately 5 mega-bases termed A and B compartments that correspond, respectively, to transcriptionally active and inactive regions in the genome (reviewed in Dekker et al. 2013; Bonev & Cavalli 2016). Oocyte maturation involves a transition from active to inactive transcription (Clegg & Piko 1983; De Leon et al. 1983), which is coupled with considerable chromatin rearrangements and condensation from a non-surrounded nucleolus to a surrounded nucleolus morphology in GV oocytes, contributing to the acquisition of developmental competence (Zuccotti et al. 1995). Single-nucleus Hi-C revealed a decline in the strength of TADs and compartments during this transition (Flyamer et al. 2017). In MII oocytes, TADs and compartments were not detected (Ke et al. 2017; Du et al. 2017) (Figure 1.7). TADs and loops start being established after fertilisation in the zygote, although with weak strength (Flyamer et al. 2017; Gassler et al. 2017) (Figure 1.7). Following the major wave of ZGA in the late two-cell embryo, TADs become more pronounced and insulation around TAD boundaries is already detected but only clearly defined from the eight-cell stage onwards (Du et al. 2017; Gassler et al. 2017; Ke et al. 2017) (Figure 1.7), suggesting a gradual shift towards a somatic-like chromatin state. This reinforcement of TADs and TAD

boundaries after ZGA has been proposed to be regulated by MERVL expression (Kruse et al. 2019). Consistently, HERVL elements also shape TADs in human ESCs (Yanxiao Zhang et al. 2019). Interestingly, genome-nuclear lamina interactions precede TAD establishment in the zygote (Borsos et al. 2019), indicating that LADs arrangement may direct higher-order chromatin topology. A and B compartments are weak and disorganised in zygotes and early two-cell embryos (Du et al. 2017; Flyamer et al. 2017; Gassler et al. 2017; Ke et al. 2017) and only detected after ZGA through to the formation of the ICM (Du et al. 2017) (Figure 1.7).

After fertilisation, the chromatin of the two parental genomes is structurally distinct. The compartment strength in the maternal genome is weaker than in the paternal genome (Du et al. 2017; Flyamer et al. 2017; Ke et al. 2017), consistent with the respective organisation seen in mature gametes, since sperm has a very structured chromatin organisation with high frequency of extra-long-range and inter-chromosome interactions (Ke et al. 2017; Jung et al. 2017). Interestingly though, TADs and loop strengths are similar between the two pronuclei (Du et al. 2017; Flyamer et al. 2017; Gassler et al. 2017; Ke et al. 2017) but fewer distal interactions are found in the paternal genome (Du et al. 2017). This allelic differences in TAD formation are regulated by cohesin and associated proteins (Gassler et al. 2017).

### 1.1.5 Interplay between epigenetic remodeling and transcriptional activation during the maternal-to-zygotic transition

All these recent transcriptional and epigenetic maps of oocytes and early embryos have shed light into the events occurring during the MZT. However, a consensus model of the functional relationships and temporal coupling between DNA demethylation, chromatin remodeling and transcriptional changes has not been established yet. For example, MERVL elements are one of the first transcripts expressed in embryos (Kigami 2002; Peaston et al. 2004; Svoboda et al. 2004; Macfarlan et al. 2012) and they also become hypomethylated following fertilisation (L. Wang et al. 2014). LINE-1, MERVL and SINE B2 elements, expressed at the time of major ZGA, show an open chromatin accessibility state and are strongly enriched in H3K4me3 at this developmental time point (Percharde et al. 2018; J. Wu et al. 2016). In fact, overall, transposable elements are present in accessible distal regions of mouse two-cell embryos and marked by H3K27ac (J. Wu et al. 2016), and, in early human embryos, SVA and HERV-K

transposons harbor distal DHS sites and are highly expressed (L. Gao et al. 2018; J. Wu et al. 2018). These data suggests that transposable elements may function not only as alternative promoters during ZGA (Peaston et al. 2004) but also as enhancer elements over long distances. Additionally, even though promoter chromatin accessibility is normally correlated with gene expression, accessible promoters are also found in inactive genes that become activated at later developmental stages, indicating these promoters are primed for activation and suggesting chromatin remodeling precedes ZGA (F. Lu et al. 2016)

These observations illustrate that whether transcriptional activity helps chromatin and DNA demethylation reprogramming or it is a consequence of these epigenetic changes remains unclear. Several studies using a combination of chemical inhibitors and gene depletion strategies have aimed to uncover such hierarchy of events. Treating zygotes or two-cell embryos with α-amanitin blocks transcription and results in developmental arrest, however, it does not prevent H3K27me3 remodeling in both parental genomes or TAD formation (H. Zheng et al. 2016; Ke et al. 2017; Du et al. 2017), indicating that chromatin reorganisation does not require ZGA. On the other hand, knock-down of the lysine-specific demehylases 5A and 5B (KDM5A and KDM5B), responsible for remodeling the broad H3K4me3 domains, resulted in downregulation of many ZGA genes (Dahl et al. 2016). Consistently, loss of H3K4me3 results in reduced paternal pronuclear ZGA (Aoshima et al. 2015) and histone hyperacetylation by trapoxin treatment enhances transcription in 2-cell embryos (Aoki et al. 1997), suggesting that chromatin changes influence transcriptional activation. However, blocking transcription in late zygotes or early 2-cell embryos impedes removal of the broad H3K4me3 domains and establishment of canonical H3K4me3 peaks (B. Zhang et al. 2016) and it also reduces chromatin accessibility and nucleosome-depleted regions at the time of ZGA (J. Wu et al. 2016; F. Guo et al. 2017), suggesting local chromatin changes require transcriptional activity. Moreover, TET3 KO or knock-down embryos undergo zygotic transcription (Inoue et al. 2012; Peat et al. 2014; Shen et al. 2014) at an increased rate (Tsukada et al. 2015), and the genes that are normally demethylated in the paternal pronucleus and then transcribed in two-cell embryos are independent on TET3-mediated demethylation (Peat et al. 2014).

Altogether, these experiments show that the epigenetic and transcriptional changes during the MZT are closely interconnected, but the exact functional interdependencies require further

investigation. Identifying the epigenetic and transcriptional regulators of this developmental transition and their mechanism of action can help towards the understanding of the precise relationships between DNA demethylation, chromatin remodeling and transcriptional activation, and this is, precisely, the aim of my dissertation.

### 1.1.6 Maternal regulators of the maternal-to-zygotic transition

The extensive chromatin remodeling and transcriptional changes that occur after fertilisation and during early embryogenesis need to be tightly regulated to ensure a correct developmental progression. Maternal effect genes (MEGs) are defined as genes whose function is essential for embryonic development and whose disruption in the oocyte before fertilisation causes developmental defects in the embryo (reviewed in K.-H. Kim & K.-A. Lee 2014; Lei Li et al. 2010). A number of them have been described in mice to date, including proteins from the subcortical maternal complex (Tong et al. 2000; Esposito et al. 2007; Yurttas et al. 2008; Lei Li et al. 2008; P. Zheng & Dean 2009; Tashiro et al. 2010; X.-J. Yu et al. 2014; Mahadevan et al. 2017); proteases such as GRANZYME G (Tsai et al. 2010); structural molecules such as β-CATENIN (de Vries et al. 2004) or FORMIN 2 (Leader et al. 2002); ubiquitylation-related enzymes (Roest et al. 2004; Sekiguchi et al. 2006; C. Yu et al. 2013; Mtango et al. 2012; Maenohara et al. 2017); cell cycle regulators such as CYCLIN A2 (Hara et al. 2005); DNA damage repair proteins such as the mismatch repair endonuclease PSM2 (Gurtu et al. 2002), breast carcinoma amplified sequence 2 (BCAS2) (Q. Xu et al. 2015) or RAD9 checkpoint clamp component A (RAD9A) (L. Huang et al. 2019); pluripotency factors such as POU5F1 (Foygel et al. 2008) or SOX2 (Pan & Schultz 2011); autophagy-related proteins such as ATG5 (TSUKAMOTO et al. 2008) or DDB1-cullin-4-associated factors 2 and 13 (DCAF2 and DCAF13) (Y.-W. Xu et al. 2017; Yang Liu et al. 2019); methionine adenosyltransferase such as MAT2A (Sun et al. 2018); and proteins involved in post-transcriptional mRNA stability and degradation (Lykke-Andersen et al. 2008; Murchison et al. 2007; Tang et al. 2007; J. Chen et al. 2011; Sousa Martins et al. 2016; Mak et al. 2018; K. Lin et al. 2018).

Several of these MEGs have been directly implicated in essential processes of the MZT, such as maternal transcript degradation and ZGA (reviewed in Lei Li et al. 2010; Eckersley-Maslin, Alda-Catalinas & Reik 2018; Vastenhouw et al. 2019). However, a vast majority of the MEGs

with direct roles in ZGA are epigenetic and transcription factors (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018), which are the subject of study of this dissertation and, therefore, I review them in detail in the following sections.

#### 1.1.6.1 DNA methylation machinery

In mouse embryos, many of the DNA methylation machinery enzymes are MEGs, consistent with the global loss of DNA methylation that occurs after fertilisation (see section 1.1.4.1) (reviewed in K.-H. Kim & K.-A. Lee 2014). These include the oocyte-specific isoform of DNMT1 (DNMT1o) (Howell et al. 2001), UHRF1 (Sharif et al. 2007; Maenohara et al. 2017), DNMT3A (Okano et al. 1999; Kaneda et al. 2004), DNMT3B (Okano et al. 1999), DNMT3L (Bourc'his et al. 2001), TET3 (T.-P. Gu et al. 2011) and GSE (Hatanaka et al. 2013).

Maternal conditional KO embryos of *Dnmt1o* and zygotic KO embryos of *Dnmt1* die during post-implantation stages due to defects in imprinting maintenance (E Li et al. 1992; Howell et al. 2001). The effect of maternal conditional KO of *Uhrf1* is more pronounced and, although oocytes present a normal phenotype, the majority of derived embryos die before the blastocyst stage with reduced levels of DNA methylation, particularly at the imprinting control regions (Maenohara et al. 2017). In contrast, zygotic KO embryos of *Uhrf1* exhibit similar developmental arrest and gestational lethality to *Dnmt1* KO embryos (Sharif et al. 2007), highlighting the importance of the maternal pools of UHRF1 to early embryogenesis. While zygotic *Dnmt3a* KO embryos survive to birth although with prominent developmental defects that result in their death a few weeks after (Okano et al. 1999), embryos derived from conditional maternal KO females die in utero and lack methylation and allele-specific expression at maternally imprinted loci (Kaneda et al. 2004). This phenotype of conditional *Dnmt3a* KO embryos is almost indistinguishable from that of *Dnmt3l* zygotic KO embryos (Bourc'his et al. 2001). The zygotic KO of *Dnmt3b* leads to demethylation at pericentric satellite DNA and severe developmental phenotypes which result in embryonic lethality from mid-gestation (Okano et al. 1999). Maternal conditional KO females for *Tet3* present reduced fecundity and their offspring show significantly compromised development (T.-P. Gu et al. 2011; Shen et al 2014; Inoue et al. 2015). Lastly, antisense RNA knock-down of GSE in

oocytes does not affect early developmental progression but derived embryos present notorious impairments in active DNA demethylation in the paternal pronucleus (Hatanaka et al. 2013).

Altogether, these studies reflect that maternal mutations in DNA methylation machinery enzymes typically cause defects at late post-implantation stages, predominantly due to imprinting misregulation, and do not normally impede ZGA, therefore suggesting that transcription in early embryos can occur even when DNA methylation is misregulated.

### 1.1.6.2 Chromatin remodellers

Eukaryotic chromatin can be remodelled by factors that use ATP hydrolysis to enhance nucleosome accessibility (reviewed in Vignali et al. 2000). This family of chromatin remodelers can be subdivided into three groups based on their biochemical properties and sequence similarity of their catalytic ATPase subunits: the SWItch/Sucrose Non-fermentable nucleosome remodeling (SWI/SNF) complex, the Imitation SWI (ISWI) complex and the Mi-2 nucleosome remodeling deacetylase (Mi-2/NuRD) complex (Boyer et al. 2000). One of the first epigenetic modifiers identified in mice as MEG and shown to activate zygotic transcription at the two-cell stage is one of the ATPase subunits of the SWI/SNF complex, namely SWI/SNF related, matrix associated, active dependent regulator of chromatin, subfamily A, member 4 (SMARCA4, also known as BRG1) (Bultman et al. 2006). Despite normal germ-cell development and fertilisation, oocyte-conditional deletions lead to embryonic arrest at the two-cell stage and impaired ZGA (Bultman et al. 2006). Interestingly, maternal mutation of the SMARCA4-interacting partner dual bromodomain and WD repeat domain containing 1 (BRWD1) results in meiotic maturation defects, and *in vitro* fertilisation (IVF) of oocytes that are able to reach the MII stage leads to developmental arrest at the pronuclear stage in zygotes (Philipps et al. 2008). These studies suggest that SMARCA4 and BRWD1 act together to regulate the MZT and highlight the importance of chromatin remodeling for efficient reprogramming to totipotency after fertilisation. Recently, the role of SMARCA4 during the MZT has also been linked to the transcriptional co-activator mediator complex 13 (MED13), whose zygotic knock-down or conditional maternal KO induces ZGA defects (Miao et al. 2018). The catalytic subunit of the ISWI complex, SMARCA5, has also been implicated in the regulation of early embryogenesis since both zygotic RNAi knock-down and KO lead to pre-

implantation developmental defects due to cell arrest in both the ICM and the trophectoderm (Stopka & Skoultchi 2011; Torres-Padilla & Zernicka-Goetz 2006), possibly due to ZGA impairments since it localises to sites of active transcription in early embryos (Torres-Padilla & Zernicka-Goetz 2006).

The polycomb group proteins are evolutionally conserved chromatin-modifying factors which canonically catalyse H3K27me3 and are critical regulators of gene expression (Schuettengruber et al. 2017). The polycomb repressive complexes 1 and 2 (PRC1 and PRC2) have roles during mouse MZT (Posfai et al. 2012; Erhardt et al. 2003). Interestingly, despite the canonical repressive function of polycomb complexes, *Ring1/Rnf2* (members of PRC1) conditional KO oocytes show defects in meiotic maturation, and derived embryos do not develop beyond the two-cell stage due to defective replication and ZGA (Posfai et al. 2012). Maternal conditional deletion of the enhancer of zeste homolog 2 (EZH2), a component of PRC2, disrupts histone H3 methylation in the early zygote and causes severe growth retardation of neonates (Erhardt et al. 2003). This late phenotype of *Ezh2*-maternal deficient embryos is likely due to embryonic rescue from the paternal allele from the four-cell stage, as complete *Ezh2*-null mice show early embryonic lethality (O'Carroll et al. 2001).

Another factor that extensively remodels the chromatin landscape in early embryos is the chaperone nucleoplasmin 2 (NPM2), whose deletion causes female infertility due to loss of heterochromatin and deacetylated histone H3 around nucleoli in oocytes and early embryos, leading to pre-implantation developmental arrest (Burns et al. 2003; La Fuente et al. 2004). The CCCTC-binding factor (CTCF) is key in the 3D organisation of the nuclear space (reviewed in Ong & Corces 2014) and, consistent with the major high-order genome reorganisation that occurs upon fertilisation (see section 1.1.4.2.3), maternal depletion of CTCF via RNAi knock-down causes meiotic defects in oocytes and mitotic defects in derived embryos that are accompanied by transcriptional activation defects during ZGA (Wan et al. 2008). Lastly, ectopic expression of the uncharacterised protein zygote arrest 1-like (ZAR1L) leads to two-cell-stage arrest, likely due to abnormal levels of H3K4me2/3 and H3K9me2/3 and downregulation of other chromatin modifiers (Hu et al. 2010), consistent with the MEG role of its closely related factor ZAR1 (Xuemei Wu et al. 2003).

#### 1.1.6.3 Histone-modifying enzymes

In recent years, the key role of a number of histone-modifying enzymes during the MZT has been uncovered (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018) (Figure 1.8). This is in fact not surprising given the extensive remodeling of histone PTMs (see section 1.1.4.2.2).

In oocytes, the lysine-lysine *N*-methyltransferase 2B (KMT2B, also known as MLL2) is the enzyme responsible for the establishment of the broad H3K4me3 domains during oogenesis (Hanna et al. 2018) and is necessary for proper ZGA (Andreu-Vieyra et al. 2010; Dahl et al. 2016) (Figure 1.8). KMT2C and KMT2D (also known as MLL3 and MLL4, respectively) were discovered as ZGA regulators through a screen of histone lysine-to-methionine (K-M) mutants essential for early embryonic development (Aoshima et al. 2015). Overexpression of H3.3 K4M and K36M mutants in MII oocytes resulted in decreased embryo survival to the blastocyst stage, whereas overexpression of these mutants in zygotes did not interfere with development, suggesting essential remodeling of these histone PTMs during the MZT. These defects are likely due to a decrease in transcription at later pronuclear stages in H3.3 K4M mutants, as shown by EU incorporation assays, indicating that remodeling of H3K4 methylation is necessary for minor ZGA (Aoshima et al. 2015). Consistently, knock-down of KMT2C/KMT2D in GV oocytes, but not in fertilised zygotes, phenocopied overexpression of H3.3 K4M mutants and resulted in decreased levels of H3K4me1 and H3K27ac (Figure 1.8), together with reduced transcription in the paternal pronucleus during the minor wave of ZGA (Aoshima et al. 2015). Regarding the phenotype of K36M mutants, the H3K36me3 methyltransferase enzyme SET domain containing 2 (SETD2) is a crucial regulator of the oocyte epigenome and its maternal deficiency results in loss of H3K36me3, failure to establish the correct DNA methylome, invasion of H3K4me3 and H3K27me3 into H3K36me3 regions, aberrant deposit of H3K4me3 at imprinting control regions and one-cell stage developmental arrest in derived embryos (Q. Xu et al. 2019). The role of SETD2 is likely regulated by the mRNA processing and export factor IWS1 that interacts with the histone H3/H4 chaperone and transcription elongation factor SUPT6 to recruit SETD2 via the AKT signalling pathway (Oqani et al. 2019). Accordingly, knock-down of both ISWI and SUPT6 in zygotes arrests embryos at the two-cell stage and leads to aberrant H3K36me3 at the time of ZGA (Oqani et

al. 2019). Recently, another H3K4 methyltransferase, the histone-lysine *N*-methyltransferase SETD1B, was reported to regulate oogenesis and early epigenetic reprogramming in the embryo, although at earlier stages than KMT2C and KMT2D (Brici et al. 2017) (Figure 1.8).

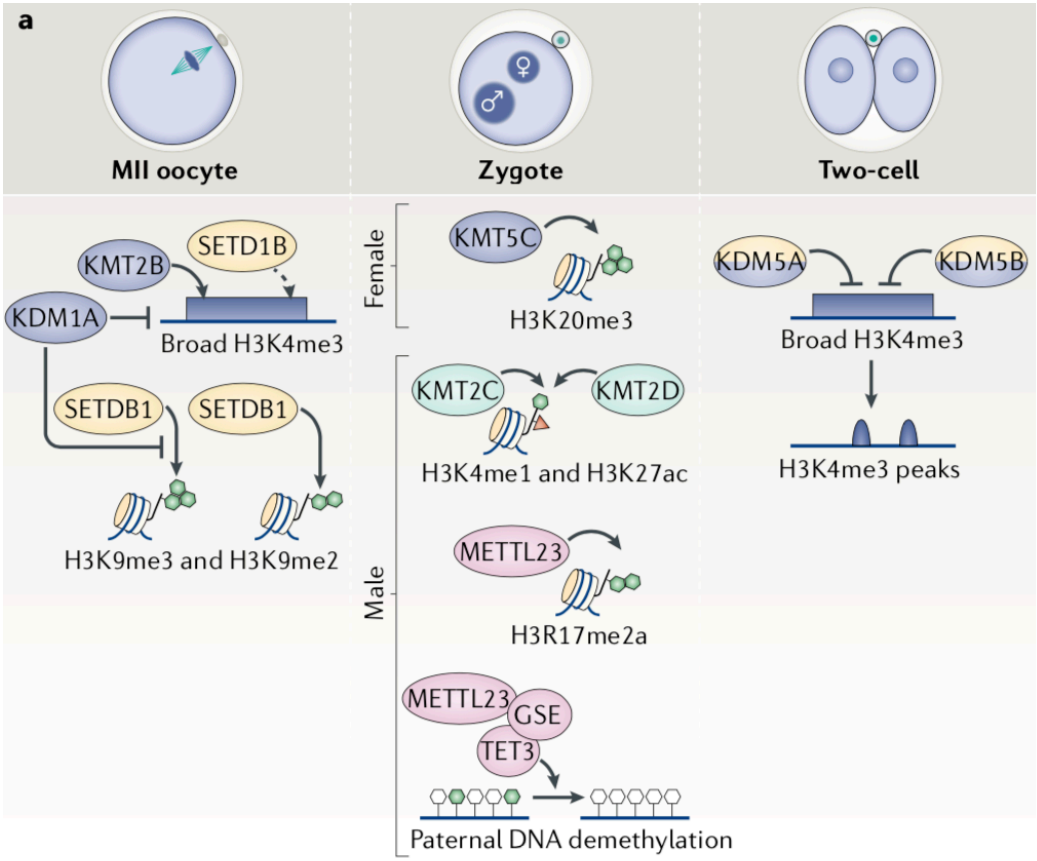



**Figure 1.8– Regulation of the MZT by histone-modifying enzymes**

Histone-lysine *N*-methyltransferase 2B (KMT2B) establishes the broad histone H3K4me3 domains in MII oocytes, which are subsequently removed by the lysine-specific demethylase 5A (KDM5A; also known as JARID1A) and KDM5B (also known as JARID1B) in two-cell embryos, leading to the formation of the canonical peaks of H3K4me3 distribution. SET domain containing 1B (SETD1B) also catalyses H3K4me3 in MII oocytes, but there is no experimental evidence that it establishes the broad H3K4me3 domains (dashed arrow). The methyltransferase SETDB1 catalyses H3K9me2 and H3K9me3 deposition in MII oocytes. H3K9me3 is removed by the H3K4me3-specific and H3K9me3-specific demethylase KDM1A (also known as LSD1). In zygotes, KMT5C catalyses H4K20me3 at the maternal pronucleus; in the paternal pronucleus, KMT2C and KMT2D catalyse the H3K4me1 and support acetylation of H3K27ac enrichment. Methyltransferase-like 23 (METTL23) catalyses the asymmetric dimethylation of histone H3 arginine 17 (H3R17me2a) and has a role in active DNA demethylation of the paternal pronucleus by recruiting the gonad-specific expression gene–ten-eleven translocation 3 (GSE–TET3) complex. Colours represent the role of the enzyme: yellow represents a role of transcription during oocyte development; green represents a role of transcription during minor ZGA; blue represents a role of transcription during major ZGA; and pink represents DNA (de)methylation proteins. Image and legend adapted from Eckersley-Maslin, Alda-Catalinas & Reik 2018.

This was observed via depletion of maternal stores of SETD1B that led to female infertility due to oocyte transcriptional misregulation, zona pellucida irregularities, meiotic defects and fertilisation impairments with polyspermy and one-cell stage arrest (Brici et al. 2017). This phenotype suggests that SETD1B might be involved in establishing and/or remodeling the broad H3K4me3 domains together with KMT2B (Figure 1.8), but further functional studies are required to confirm this hypothesis.

Following fertilisation, histone methyltransferases also play a critical role during the protamine-histone exchange of the paternal pronucleus and during heterochromatin formation. The enzyme methyltransferase-like 23 (METTL23) catalyses asymmetric dimethylation of histone H3 arginine 17 (H3R17me2a), a histone modification that together with the histone cell cycle regulator (HIRA) has been shown to be critical for incorporation of the histone variant H3.3 in the male pronucleus after fertilisation (Santenard et al. 2010; Akiyama et al. 2011; Inoue & Yi Zhang 2014; C.-J. Lin et al. 2014; Nashun et al. 2015; Hatanaka et al. 2017). Furthermore, METTL23 also acts during the wave of active DNA methylation of the paternal pronucleus by recruiting the GSE-TET3 complex (Hatanaka et al. 2017) (Figure 1.8). Regarding heterochromatin formation, the constitutive heterochromatin mark H4K20me3 is specifically found after fertilisation in the maternal pronucleus, localizing around the ring-like structures of NPBs (Probst et al. 2007; Wongtawan et al. 2011), and it is catalysed by KMT5C (also known as SUV420H2) (Figure 1.8), amongst other enzymes. Accordingly, overexpression of KMT5C in zygotes increases H4K20me3 accumulation in the maternal pronucleus and reduces transcriptional activity during ZGA, leading to developmental arrest at the one- to two-cell stages (Eid et al. 2016). H3K9 methylation silences transposable elements in pericentric repeats (Bannister et al. 2001) and is specifically enriched in the maternal pronucleus (Santos et al. 2005; Hajkova et al. 2010; Nakamura et al. 2012). Maternal KO of the methyltransferase *Setdb1*, which catalyses H3K9me2/3 (Figure 1.8) and protects the maternal pronucleus from active DNA demethylation (T.-B. Zeng et al. 2019), induces meiotic impairments, DNA damage, derepression of retrotransposons and degeneration of derived embryos before the morula stage (J. Kim et al. 2016). Similarly, maternal deletion of EHMT2 (also known as G9a), which together with SETDB1 is also responsible for the H3K9me2

asymmetry between pronuclei after fertilisation (T.-B. Zeng et al. 2019), reduces the developmental rate to blastocysts (Au Yeung et al. 2019).

Histone demethylases are also crucial in remodeling histone PTMs during the MZT. The lysine-specific demethylases responsible for removing the broad H3K4me3 domains established in oocytes and remodel them to a canonical peak structure after fertilisation are KDM5A and KDM5B (also known as JARID1A and JARID1B, respectively) (Dahl et al. 2016; Xiaoyu Liu et al. 2016; B. Zhang et al. 2016) (Figure 1.8). Consequently, ectopic expression of KDM5B, but not KDM5A, leads to reduced levels of H3K4me3 and reactivation of transcription in surrounded nucleolus GV oocytes (B. Zhang et al. 2016), whereas KDM5A and KDM5B morpholino-mediated knock-down in zygotes downregulates the expression of many major ZGA genes at the two-cell stage and impairs pre-implantation development (Dahl et al. 2016; B. Zhang et al. 2016), likely due to failure to remodel the broad H3K4me3 domains. Therefore, KDM5A and KDM5B have a role both in oocyte transcription and ZGA (Figure 1.8). Another lysine demethylase critical for ZGA is KDM1A (also known as LSD1), whose depletion in maternal stores arrests derived embryos at the one- to two-cell stages (Ancelin et al. 2016; Wasson et al. 2016). This demethylase acts on H3K4me3 and H3K9me3 (Figure 1.8) and, in agreement, two-cell embryos derived from KDM1A maternal mutants show increased levels of these histone PTMs, increased amounts of maternal transcripts and LINE-1 elements and fail to undergo ZGA (Ancelin et al. 2016; Wasson et al. 2016). This phenotype suggests that KDM1A helps remodeling the broad H3K4me3 domains and H3K9me3 upon fertilisation (Figure 1.8).

The repressive histone modification H3K27me3 is regulated by the H3K27me2/3-specific demethylases KDM6A and KDM6B (also known as UTX and JMJD3, respectively), whose functions are contrasting during pre-implantation development (L. Yang et al. 2016). For example, KDM6B knock-down in MII oocytes improves the developmental efficiency of parthenogenetic embryos, whereas KDM6A knock-down compromises development to the blastocyst stage; double maternal knock-down of KDM6A and KDM6B significantly reduces developmental rates (L. Yang et al. 2016). Consistent with these observations, ectopic expression of KDM6A improves the efficiency of somatic cell nuclear transfer (SCNT) whereas knock-down of KDM6B facilitates ZGA in these embryos, as measured with a

MERVL reporter (L. Yang et al. 2018). These studies suggest that KDM6A and KDM6B demethylases are tightly regulated to remodel H3K27me3 during the MZT, although the exact mechanism of action remains unknown.

### 1.1.6.4 Transcription factors and DNA binding proteins

During ZGA, a number of maternal transcription factors bind gene promoters and activate transcription in the zygote and two-cell embryo. Amongst the first identified is the transcription intermediary factor 1α (TIF1α), that translocates to the nucleus during mid-to-late pronuclear stages in the zygote and localises with other maternal chromatin remodellers, such as SMARCA4 and SMARCA5, to regulate RNA polymerase II localisation and transcription in the zygote, as shown by staining of 5-bromouridine-5′-triphosphate incorporation (Torres-Padilla & Zernicka-Goetz 2006).

Another recently discovered critical regulator is the nuclear transcription factor Y subunit-α (NFYA), that was identified via motif enrichment analysis of open DHS sites in two-cell embryos (F. Lu et al. 2016) and its role is conserved in zebrafish (Stanney et al. 2019). NFYA partly contributes to ZGA given that knock-down in GV oocytes leads to downregulation of approximately 15% of two-cell activated genes in derived embryos, as measured by RNA-sequencing, but also results in embryonic arrest at the morula stage and reduced chromatin openness at gene promoters in two-cell embryos (F. Lu et al. 2016).

STELLA is a key regulator of the oocyte DNA methylation landscape and it protects against active DNA demethylation in the maternal pronucleus after fertilisation (Hajkova et al. 2010; Nakamura et al. 2012; Y. Li et al. 2018). Interestingly, several studies also show its role as a MEG (Payer et al. 2003; Bortvin et al. 2004) by regulating chromatin remodeling and major ZGA (Arakawa et al. 2015; Y. Huang et al. 2017; Y. Li et al. 2018). *Stella*-null oocytes show excessive levels of DNA methylation genome-wide, which is inherited by the zygote and maintained in two-cell embryos (Y. Li et al. 2018). Additionally, *Stella*$^{-/-}$ embryos show impaired chromocenter formation due to failure to undergo reverse transcription of major satellites and H3.3 incorporation (Arakawa et al. 2015). In fact, death-associated protein 6 (DAXX), responsible for H3.3 incorporation in pericentromeric regions, is reduced in *Stella*$^{-/-}$

zygotes causing these defects, as demonstrated by rescue experiments (Arakawa et al. 2015). Recent single-cell and single-embryo RNA-sequencing studies suggest that this phenotype is linked to failure to transcribe ZGA genes, including MERVL elements, and to downregulate maternal transcripts (Y. Huang et al. 2017; Y. Li et al. 2018).

Highlighting the importance of tight regulation of degradation of maternal transcripts concomitantly with ZGA was the recent discovery of the transcriptional co-activator yes-associated protein 1 (YAP1) as a critical regulator of the MZT (C. Yu, Ji, Dang, et al. 2016). YAP1 protein accumulates in oocytes and contributes to the degradation of up to approximately 80% of maternal transcripts and to the activation of around 700 ZGA genes (C. Yu, Ji, Dang, et al. 2016). Consequently, depletion of maternal stores leads to a prolonged two- to four-cell stage transition and reduced pre-implantation developmental rates (C. Yu, Ji, Dang, et al. 2016).

The homologs *Dux* (mouse) and *DUX4* (human) multicopy retrogenes encode double-homeodomain proteins and are conserved throughout placental mammals (Leidenroth et al. 2012). Three companion studies recently uncovered a role for DUX/DUX4 during ZGA in both species (De Iaco et al. 2017; Hendrickson et al. 2017; Whiddon et al. 2017). Dux/Dux4 transcripts are detected at the onset of zygotic transcription (De Iaco et al. 2017; Hendrickson et al. 2017) (Figure 1.9) and their identification came through two different observations: 1) DUX4/DUX activates an early embryonic two-cell-like gene signature in human facioscapulohumeral muscular dystrophy cells (FSHD) and in mouse myoblast, respectively (De Iaco et al. 2017; Hendrickson et al. 2017; Whiddon et al. 2017) (Figure 1.9); and 2) motif analysis of cleavage-stage specific genes, as defined from transcriptome analysis of human oocytes and pre-implantation embryos, showed enrichment of DUX4 motif binding in the TSS (Hendrickson et al. 2017). Furthermore, in mouse and human ESCs, DUX induces activation of ZGA-like transcripts and retrotransposons, including MERVL elements (De Iaco et al. 2017; Hendrickson et al. 2017) (Figure 1.9). ChIP-seq analysis showed that up to 20% of DUX-induced genes and ~50% of MERVL elements are directly bound by DUX (Hendrickson et al. 2017). Strikingly, DUX is a potent inductor of the rare two-cell-like (2C-like) subpopulation within mouse ESCs (mESCs) (see section 1.2), as analysed using a MERVL LTR reporter (De Iaco et al. 2017; Hendrickson et al. 2017) (Figure 1.9). Additionally, transcriptional analysis in

two-cell embryos showed that absence of zygotic DUX leads to failure to activate ZGA-specific targets, including MERVL elements (De Iaco et al. 2017).

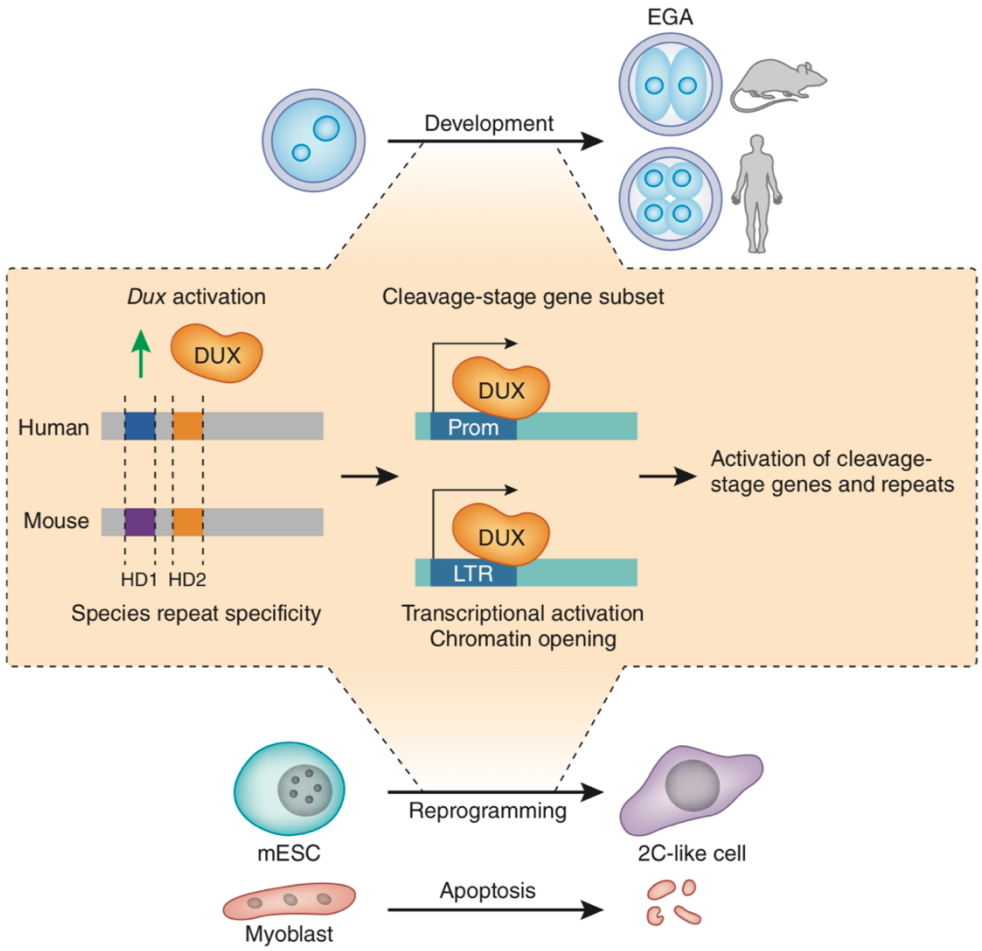



**Figure 1.9– DUX/DUX4 as a regulator ZGA**

DUX/DUX4 is a multicopy retrogene, conserved in human and mouse, that regulates ZGA or embryonic genome activation (EGA) at the two-cell stage in mouse and at the four-cell stage in human. It is first expressed in the zygote or one-cell embryo, where it binds to promoters (Prom) and LTR elements associated to ZGA or cleavage-stage genes, leading to chromatin opening and transcriptional activation of these transcripts. *In vitro*, it induces the 2C-like subpopulation in mouse embryonic stem cells (mESCs) and leads to apoptosis in both mouse and human myoblasts. Image from Iturbide & Torres-Padilla 2017.

Interestingly, despite these exciting results that suggested that DUX was the master regulator of ZGA, recent studies have shown that *Dux* maternal and zygotic KO mouse embryos can survive after birth (Iaco et al. 2019; Z. Chen & Yi Zhang 2019), and DUX4 siRNA-mediated know-down in human embryos leads to transcriptome miregulation but embryos do not arrest

(Vuoristo et al. 2019). These latest studies indicate that DUX is a non-essential synchronizer of ZGA. Consistently, DUX/DUX4 have not been detected as proteins in the oocyte, which suggests that there should be other maternal stores in the oocyte prior to fertilisation that allow Dux transcription during the minor wave of ZGA.

Maternal regulators of ZGA are less well characterised in human embryos due to the ethical restrictions to manipulate them. Besides DUX4, another transcription factor that was recently identified by motif enrichment analysis of gene promoters activated during human ZGA is pleomorphic adenoma gene 1 (PLAG1), which binds to SINE Alu elements in human and SINE B1 elements in mouse (Madissoon et al. 2019). Maternal deletion of PLAG1 in mouse oocytes leads to transcriptional misregulation and an extended two-cell stage (Madissoon et al. 2019). Motif analysis of open regions in chromatin accessibility maps can also help shed light into the potential regulators of human ZGA (J. Wu et al. 2016; Lin Li et al. 2018; L. Gao et al. 2018). In these regards, the pluripotency factor OCT4 is enriched in DHS sites at the time of ZGA in humans, but not in mice, and consistently, siRNA knock-down in human, but not in mouse, zygotes leads to downregulation of ZGA genes at the eight-cell stage (L. Gao et al. 2018). Other candidate factors implicated in human ZGA are PRD-like homeobox genes (Töhönen et al. 2015; Jouhilahti et al. 2016), but their functional role remains to be probed. Recently, Krüppel-like factors (KLFs) have recently been shown to activate the expression of evolutionally-recent transposable elements during human ZGA, whose expression remodels the chromatin landscape and is subsequently repressed by Krüppel-associated box (KRAB)-containing zinc finger proteins (KZFPs) in a timely manner (Pontis et al. 2019).

Notably, none of these transcription factors seem to be the master regulator of mouse ZGA. While oocytes accumulate STELLA and YAP1 proteins that are inherited after fertilisation (Payer et al. 2003; Bortvin et al. 2004; C. Yu, Ji, Dang, et al. 2016), they only activate a subset of ZGA transcripts (C. Yu, Ji, Dang, et al. 2016; Y. Huang et al. 2017; Y. Li et al. 2018). Although promising to be a pioneer transcription factor during ZGA, *Dux*/*DUX4* is only expressed from the zygotic genome (De Iaco et al. 2017; Hendrickson et al. 2017) (Figure 1.9), mouse KO females have viable offspring (Iaco et al. 2019; Z. Chen & Yi Zhang 2019) and human knock-down embryos do not arrest (Vuoristo et al. 2019). Therefore, all these studies suggest that ZGA is regulated by a complex network of redundant regulatory loops and factors.

Consequently, systematic studies, such as the one presented in this dissertation, are needed to improve our understanding of ZGA and its regulation. Recently, using the *in vitro* 2C-like mESC subpopulation, three independent studies, including one described in chapter 5 of this dissertation, identified the Developmental Pluripotency-Associated proteins 2 and 4 (DPPA2 and DPPA4) as critical regulators of ZGA via regulation of the expression of Dux (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Y.-L. Yan et al. 2019) (see section 1.2.3). Although DPPA2 overexpressing mESCs show contribution to extraembryonic tissues in chimeric embryos, highlighting the increased developmental potential of these cells (Y.-L. Yan et al. 2019), functional *in vivo* studies of DPPA2 and DPPA4 are still missing. Maternal depletion of these proteins and analysis of transcription in derived embryos is needed to assess how critical they are in regulating ZGA.

#### 1.1.6.5 Non-coding RNAs

Importantly, not only proteins but also non-coding RNAs play a role in the MZT (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018; Svoboda 2017) (Figure 1.10). The microRNA family miR-125 is a negative regulator of ZGA that controls the expression of the maternal-effect genes homeobox SEBOX and Lin-28 homolog A (LIN28A) (K.-H. Kim et al. 2016) (Figure 1.10). Injection of miR-125 family member mimics in GV oocytes causes embryonic arrest at the two-cell stage while inhibitors of this microRNA family enhance ZGA (K.-H. Kim et al. 2016).

Interestingly, parental diet is important for MERVL activation and ZGA. This was demonstrated by a study showing that protein restriction increases the levels of GCC 5' fragments of glycine tRNAs in mouse sperm, which are inherited into the embryo after fertilisation and repress MERVL elements during the major wave of activation in two-cell embryos (Sharma et al. 2016) (Figure 1.10).

As described in section 1.1.3.3, LINE-1 elements regulate chromatin accessibility in early embryos (Figure 1.10), since prolonging its activation beyond the two-cell stage leads to an increase in DNaseI sensitivity and, in agreement, premature silencing reduces it (Jachowicz et al. 2017). Consistently, sustained expression of LINE-1 elements beyond the two-cell stage, as

well as their repression immediately after fertilisation, impairs development to the blastocyst (Beraldi et al. 2006; Jachowicz et al. 2017), suggesting that this tight window of LINE-1 expression is critical for proper development. Interestingly, injection of the LINE-1 transcript in two-cell embryos has no effect on blastocyst development, indicating that it is the act of LINE-1 transcription itself rather than the products of transcription that regulate chromatin remodeling (Jachowicz et al. 2017). LINE-1 expression is linked to *Dux* silencing and exit from the two-cell stage (Percharde et al. 2018). In the *in vitro* 2C-like subpopulation, LINE-1 transcripts provide a nuclear RNA scaffold that recruits Nucleolin and KAP1, subsequently repressing *Dux* (Percharde et al. 2018), however, whether this mechanism also applies to early embryos remains unknown.

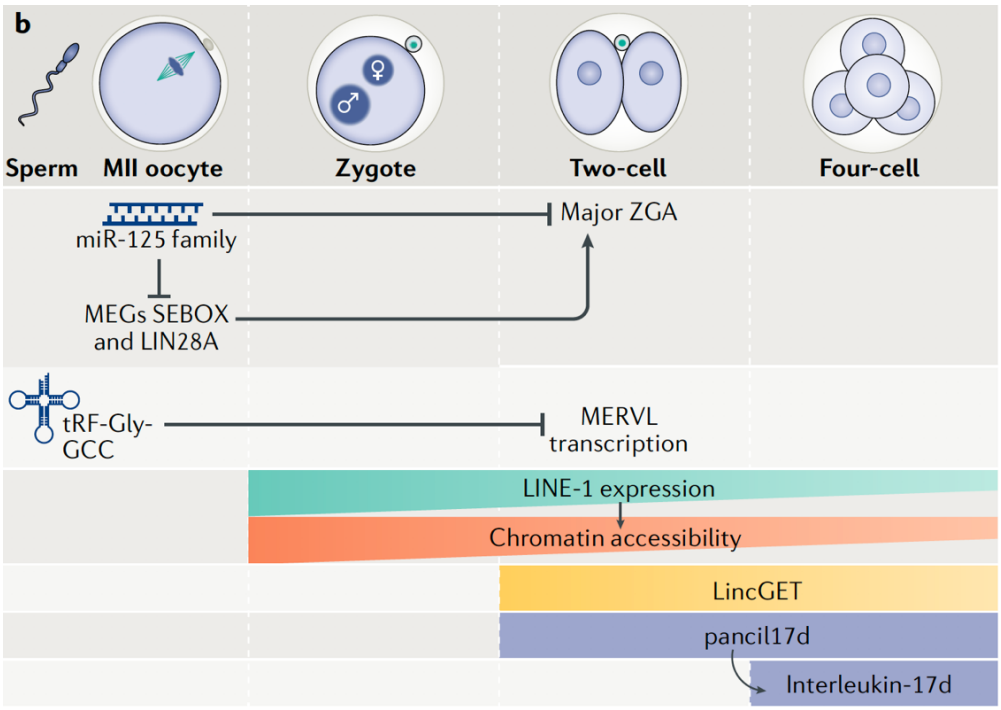



**Figure 1.10– Regulation of the MZT by non-coding RNAs**

Members of the microRNA miR-125 family are expressed in MII oocytes and suppress the expression of maternal-effect genes (MEGs), including SEBOX and protein lin-28 homologue A (LIN28A), consequently repressing transcription during the major wave of ZGA in two-cell embryos. The 5′ fragments of tRNA-Gly-GCC (tRF- Gly-GCC) are inherited by the zygote from the sperm under certain dietary conditions and negatively regulate the expression of MERVL in two-cell embryos. LINE-1 repeats are expressed from the zygote up to four- cell embryos and regulate chromatin accessibility. The long non-coding RNA LincGET is expressed in two-cell and four-cell embryos, similarly to promoter-associated non-coding RNA with interleukin-17d (pancIl17d), which promotes the expression of interleukin-17d from the four-cell stage. Image and legend adapted from Eckersley-Maslin, Alda-Catalinas & Reik 2018.

Lastly, the long non-coding RNA (lncRNAs) LincGET is expressed in two-cell and four-cell embryos (Figure 1.10) and its depletion compromises development beyond the two-cell stage, likely due to impairments in RNA splicing and MAP kinase (MAPK) signalling (Jiaqiang Wang et al. 2016). The associated non-coding RNA with interleukin-17d (pancIl17d) is a lncRNA highly expressed at the time of ZGA that promoters Il17-d expression in four-cell embryos (Hamazaki et al. 2015) (Figure 1.10); its knock-down in zygotes reduces the developmental rate to blastocysts through apoptosis-mediated mechanisms (Hamazaki et al. 2015).

## 1.2 Mouse embryonic stem cells to study zygotic genome activation

Mouse embryonic stem cells (mESCs) can be isolated from the ICM of early blastocysts and cultured *in vitro*, serving as an invaluable tool for mechanistic studies in developmental biology (M. J. Evans & Kaufman 1981; G. R. Martin 1981). When cultured in serum-containing media with Leukaemia Inhibitory Factor (LIF), mESC have self-renewal capacity and are pluripotent (Figure 1.11). Under these culture conditions, mESCs are heterogeneous both at the transcriptional and epigenetic level (reviewed in De Los Angeles et al. 2015; H. J. Lee et al. 2014; Torres-Padilla & Chambers 2014). Recently, a rare and transient sub-population within mESC was identified to have totipotent-like properties (Morgani et al. 2013; Macfarlan et al. 2012; Choi et al. 2017; Y.-L. Yan et al. 2019) and express transcripts normally restricted to two-cell stage embryos, such as MERVL (Macfarlan et al. 2012) and the zinc finger and SCAN domain containing-4 (ZSCAN4) gene cluster (Zalzman et al. 2010), amongst others (Bošković et al. 2014; Ishiuchi et al. 2015; Akiyama et al. 2015; Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018) (Figure 1.11). These cells are termed two-cell-like cells (2C-like cells) and represent an *in vitro* approach to mimic totipotency and ZGA, both from a transcriptional and epigenetic point of view (Macfarlan et al. 2012; Zalzman et al. 2010; Bošković et al. 2014; Ishiuchi et al. 2015; Akiyama et al. 2015; Eckersley-Maslin et al. 2016; Dan et al. 2017; Rodriguez-Terrones et al. 2018; reviewed in Ishiuchi & Torres-Padilla 2013) (Figure 1.11).

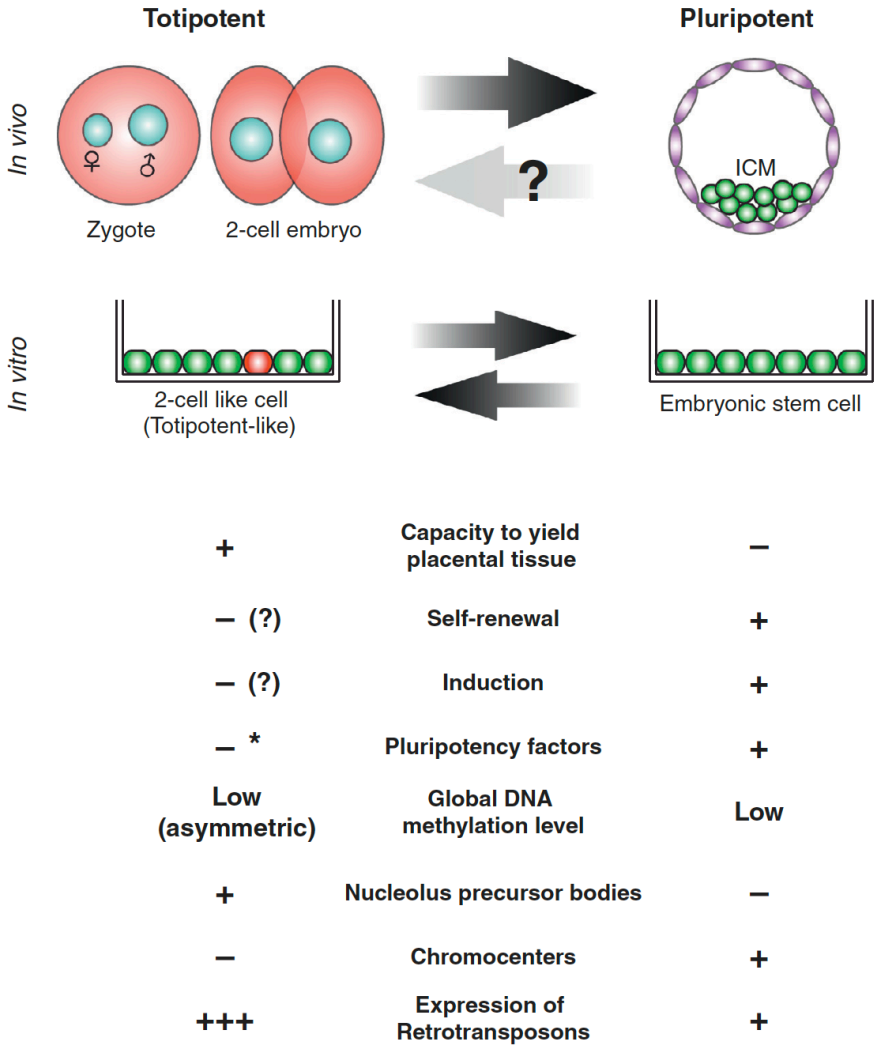



**Figure 1.11– Features of totipotent-like and pluripotent mESCs**

2-cell (2C)-like cells can be used as an *in vitro* approximation of the 2-cell embryo and are a transient sub-population naturally present in mESC cultures, which derived from the inner cell mass (ICM) of blastocysts. While mESCs are pluripotent, 2C-like cells present totipotent-like features reminiscent of zygotes and 2-cell embryos, such as the capacity to yield placental tissue. Pluripotent mESCs have the capacity to self-renew and a pluripotent state can be induced experimentally; however, 2C-like cells in culture cycle in and out of this totipotent-like state and whether totipotency can be induced *in vitro* remains controversial. Some pluripotency factors, such as OCT4, are not present in 2C-like cells, although the transcript is expressed (*). Pre-implantation mouse embryos are characterised by low global DNA methylation levels, which are asymmetric in the zygote and two-cell embryo according to the parental origin; *in vitro,* cycling through the 2C-like state involves DNA demethylation, which is restored after the cells exit this transient state, except in imprinted regions. Nucleolus precursor bodies are observed in zygotes and 2-cell embryos but these disappear after the 4-cell stage and are replaced by chromocenters, which are observed in mESCs but not in the 2C-like sub-population. Retrotransposons are particularly expressed in early embryos after fertilisation and in 2C-like cells. Arrows indicate the transitions between the two stages: zygote and two-cell embryos develop into blastocysts and 2C-like cells within mESC cultures cycle in and out of this state. Image from Ishiuchi & Torres-Padilla 2013.

### 1.2.1 Transcriptional features of 2C-like cells: MERVL and Zscan4

2C-like cells can be identified in culture with reporters for MERVL and ZSCAN4 (Macfarlan et al. 2012; Zalzman et al. 2010; Ishiuchi et al. 2015; Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018). MERVL is an LTR retrotransposon described in detail in section 1.1.3.3 of this introduction. Approximately 1% of mESC express MERVL (Macfarlan et al. 2012) (Figure 1.11). In these cells, the MERVL MT2 promoter drives expression of hundreds of chimeric uncharacterised transcripts that are organised into tight genomic clusters (Macfarlan et al. 2012; Eckersley-Maslin et al. 2016; Choi et al. 2017; Rodriguez-Terrones et al. 2018), suggesting that MERVL induces coordinated and rapid regulation of these genes. In fact, 2C-like specific genes tend to be close to MERVL MT2 elements (Eckersley-Maslin et al. 2016; X. Fu et al. 2019).

The *Zscan4* gene cluster includes six genes (*Zscan4a, Zscan4b, Zscan4c, Zscan4d, Zscan4e* and *Zscan4f)* and three pseudogenes (*Zscan4-ps1, Zscan4-ps2* and *Zscan4-ps3*) whose expression peaks at the late 2C-stage in mouse embryos (Falco et al. 2007; Ishiguro, Nakatake, et al. 2016), although recent reports suggest they are also expressed in adult testes and ovaries (Ishiguro, Nakatake, et al. 2016; Ishiguro, Monti, et al. 2016). Around 5% of mESCs express ZSCAN4 at a given time (Zalzman et al. 2010; Falco et al. 2007). Transient cycling of mESCs through a ZSCAN4$^+$ state has been shown to promote telomere extension, accounting for a role of this gene cluster in genomic stability (Zalzman et al. 2010) and maintenance of developmental potency in long-term cultures (Hirata et al. 2012; Amano et al. 2013).

Approximately 1-2% of mESCs express both MERVL and ZSCAN4 (Eckersley-Maslin et al. 2016) together with hundreds of genes expressed in the two-cell embryo at the time of ZGA (Macfarlan et al. 2012; Bošković et al. 2014; Ishiuchi et al. 2015; Akiyama et al. 2015; Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018). Importantly, the 2C-like sub-population labelled by these two reporters exist in all stages of the cell cycle, similar to mESCs, although with a prolonged G2/M phase (Storm et al. 2014, Eckersley-Maslin et al. 2016). While all MERVL$^+$ cells present high expression of ZSCAN4, not all ZSCAN4$^+$ cells express MERVL (Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018). Single-cell transcriptomics analysis suggest that this single ZSCAN4$^+$ population likely represents an

intermediate stage in the induction of two-cell embryo transcripts (Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018). Nevertheless, the functional relationship between MERVL and ZSCAN4 expression remains unclear. Recently, ChIP-seq analysis showed that ZSCAN4 binds to enhancers and MT2 elements, recruits chromatin remodellers such as BRG1 and, consequently, enhances MERVL expression and associated ZGA transcripts by promoting H3K4me1, H3K27ac and H3K14ac deposition on MT2 (W. Zhang et al. 2019). The role of ZSCAN4 in enhancing rather than directly regulating ZGA-like expression was also previously proposed in a model in which DPPA2 and DPPA4 regulate DUX expression, which in turns activates ZSCAN4 that helps enhancing the 2C-like state (Eckersley-Maslin et al. 2019).

Furthermore, pluripotency proteins such as OCT4 or SOX2 are absent in 2C-like cells, whereas the mRNA transcripts are still present (Macfarlan et al. 2012; Ishiuchi et al. 2015; Eckersley-Maslin et al. 2016; Choi et al. 2017; Rodriguez-Terrones et al. 2018) (Figure 1.11). Notably, this downregulation of pluripotency proteins is unrelated to differentiation (Rodriguez-Terrones et al. 2018).

Interestingly, both MERVL and ZSCAN4 transiently activate the expression of genes expressed in the two-cell embryo during intermediate-to-late stages of induced pluripotent stem cell reprogramming and its overexpression can increase reprogramming efficiency (Hirata et al. 2012; Jiang et al. 2013; Friedli et al. 2014; Y.-W. Kwon et al. 2015; Eckersley-Maslin et al. 2016; T. Zhao et al. 2018), suggesting that cellular reprogramming to pluripotency requires transient passage through a totipotent-like state.

### 1.2.2 Epigenetic features of 2C-like cells

Importantly, in addition to the transcriptome, the chromatin and the DNA methylation landscape of 2C-like cells also resembles that of the two-cell mouse embryo (Figure 1.11). Similar to the protein downregulation observed for pluripotency factors, the DNA methylation machinery, such as DNMT1, UHRF1, DNMT3A and DNMT3B is downregulated at the protein level in 2C-like cells (Eckersley-Maslin et al. 2016; Dan et al. 2017), leading to a global loss of DNA methylation genome-wide reminiscent to the early embryo, as measured by mass spectrometry and bisulfite sequencing (Eckersley-Maslin et al. 2016) (Figure 1.11). This

coincides with an increase in 5-hydroxymethylation (Eckersley-Maslin et al. 2016). Strikingly, after exiting the transient 2C-like state, cells recover the expression of DNA methyltransferases and restore their DNA methylation levels, except at imprinted regions (Eckersley-Maslin et al. 2016) (Figure 1.11). Mechanistically, ZSCAN4 proteins expressed in 2C-like cells recruit the DNA methylation maintenance proteins UHRF1 and DNMT1 and target their degradation through the E3 ubiquitin ligase activity of UHRF1 (Dan et al. 2017). Importantly, DNA methylation analysis of *Dnmt1* KO mESCs revealed that the entry in the 2C-like state and expression of MERVL- and ZSCAN4-regulated transcripts is not a consequence of the global DNA demethylation but rather DNA demethylation occurs subsequently to MERVL/ZSCAN4-transcriptional network activation (Eckersley-Maslin et al. 2016).

ATAC-seq analysis of the 2C-like subpopulation showed that these cells, similar to the two-cell embryo, have a globally decondensed chromatin, with accessibility especially pronounced at promoters of ZGA-like genes and MERVL elements (Eckersley-Maslin et al. 2016). Consistently, these cells lack chromocenters (Ishiuchi et al. 2015) (Figure 1.11) and their chromatin mobility, as measured by fluorescent recovery after photobleaching (FRAP) of the histones H3.1 and H2A, is higher than in the bulk of mESCs and similar to that observed in two-cell embryos (Bošković et al. 2014; Ishiuchi et al. 2015). This can be further induced by downregulation of the Chromatin-Assembly Factor 1 (CAF-1), which increases the 2C-like subpopulation (see section 1.2.3, Figure 1.11) and leads to higher H3.1 mobility and dispersed chromocenters (Ishiuchi et al. 2015). In agreement with an increased chromatin accessibility, histone acetylation is also increased in 2C-like cells (Macfarlan et al. 2012; Ishiuchi et al. 2015; Akiyama et al. 2015), leading to derepression of heterochromatin and transcriptional activation of major satellite repeats in pericentromeric chromatin (Ishiuchi et al. 2015; Akiyama et al. 2015), a feature reminiscent of zygotes and two-cell embryos (Probst et al. 2010; Santenard et al. 2010). In fact, treating mESCs with the histone deacetylase inhibitor sodium butyrate induces the 2C-like subpopulation (Dan et al. 2015). At the higher-order chromatin structure, recent Hi-C analysis of CAF-1 knock-down-induced 2C-like cells have shown that these cells present an increased 3D structural plasticity, with less long-range contacts, reduced strength in TADs and loops and an overall decreased compartment strength (Kruse et al. 2019).

### 1.2.3 Regulators of 2C-like cells

Given the transcriptional and epigenetic similarities between 2C-like mESCs and two-cell embryos, these cells can be used as *proxy* to mimic ZGA *in vitro*. Consequently, a number of studies have aimed to identify factors which, upon deletion or overexpression, increase the 2C-like subpopulation (Figure 1.12).

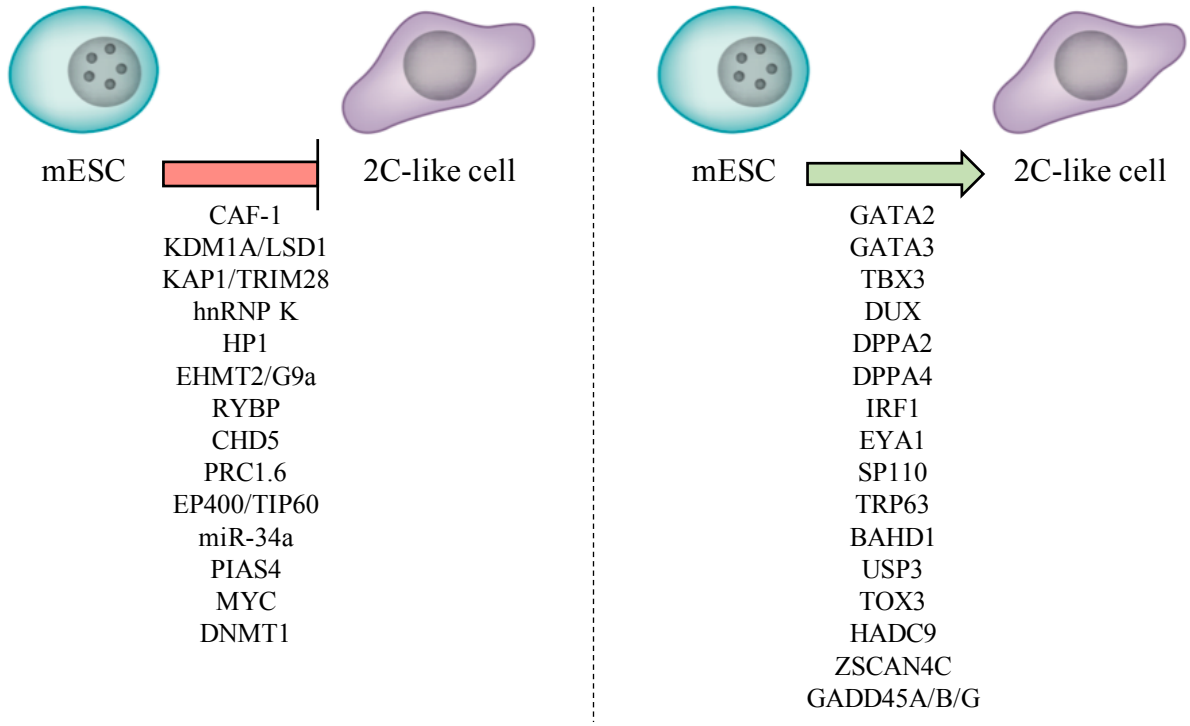



**Figure 1.12– Regulators of 2C-like cells**

Identified repressors (left) and inducers (right) of MERVL expression and/or the 2C-like subpopulation in mESCs. CAF-1: chromatin-assembly factor 1 (CAF-1); KDM1A (also known as LSD1): lysine demethylase 1A; KAP1 (also known as TRIM28): KRAB-associated protein-1; hnRNP K: heterogeneous nuclear ribonucleoprotein K; HP1: heterochromatin protein-1; EHMT2 (also known as G9a): histone-lysine N-methyltransferase 2; RYBP: RING1 and YY1-binding protein; CHD5: chromodomain helicase DNA binding protein-5; PRC1.6: polycomb complex 1.6; EP400 (also known as TIP60): E1A-binding protein p400; PIAS4: SUMO2 E3 ligase; MYC: MYC proto-oncogene, BHLH transcription factor; DNMT1: DNA methyltransferase-1; GATA2/3: GATA-binding factor 2/3; TBX3: T-box transcripton factor-3; DUX: double homeobox; DPPA2/4: developmental pluripotency-associated proteins 2/4; IRF1: interferon regulator factor-1; EYA1: transcriptional coactivator and phosphatase-1; SP110: SP110 nuclear body protein; TRP63: tumour protein 63; BAHD1: bromo adjacent homology domain containing-1; USP3: ubiquitin specific peptidase-3; TOX3: TOX high mobility group box family member-3 (TOX3); HDAC9: histone deacetylase-9 (HDAC9); ZSCAN4C: zinc finger and SCAN domain containing-4C. GADD45A/B/G: growth arrest and DNA damage-inducible proteins-45A/B/G. Cell schematics from Iturbide & Torres-Padilla 2017.

As mentioned above, interfering with epigenetic factors, such CAF-1, can increase the 2C-like subpopulation (Figure 1.12), as labelled by a MERVL reporter, 10- to 30-fold with a subsequent increase of ZGA-like transcripts, including MERVL and Zscan4 (Ishiuchi et al. 2015). Interestingly, SCNT embryos derived from CAF-1 knock-down mESCs showed higher reprogrammability, suggesting an increased developmental potency of these cells (Ishiuchi et al. 2015). Knock-down of other repressive chromatin complexes also results in an expansion of MERVL-expressing mESCs (reviewed in Schlesinger & Goff 2015). For example, KDM1A (also known as LSD1), whose function is critical in the early embryo (Ancelin et al. 2016; Wasson et al. 2016), is a key repressor of endogenous retroviruses in mESCs, specifically, it regulates a subset of ZGA genes whose expression is controlled by a MERVL LTR promoter (Macfarlan et al. 2011) (Figure 1.12). This function of KDM1A is likely exerted via chromatin regulation of a number of repressive histone PTMs, such as hypomethylation of H3K4, hypoacetylation of H3K27 and dimethylation of H3K9 at ERVs (Macfarlan et al. 2011). Consequently, *Kdm1a* mutant mESCs express high levels of ZGA genes and show higher contribution to extra-embryonic tissues in embryoid-body differentiation assays (Macfarlan et al. 2011). KAP1 (also known as TRIM28), is a repressor that acts by recruiting SETDB1, heterochromatin protein-1 (HP1) and the NuRD complex (Sripathy et al. 2006) (Figure 1.12). Similar to KDM1A, *Kap1* deletion in mESCs leads to upregulation of ERVs, specifically intracisternal A-type particles (IAPs) and MERVL elements, possibly due to misregulation of H3K9 methylation and H4 acetylation (Rowe et al. 2010; Maksakova et al. 2013). This phenotype is similar to that observed in KO cells of SETDB1/KAP1 binding partner Heterogeneous Nuclear Ribonucleoprotein K (hnRNP K) (Thompson et al. 2015) (Figure 1.12). Deletion of HP1 and the H3K9 methyltransferase EHMT2 or G9a, also leads to deregulation of MERVL elements and associated ZGA genes (Maksakova et al. 2013) (Figure 1.12), consistent with their role in the early embryo (T.-B. Zeng et al. 2019; Au Yeung et al. 2019). The PRC1 protein RING1 has functions during meiotic maturation, mitotic cleavage divisions and ZGA (Posfai et al. 2012). *In vitro*, the RING1 and YY1-binding protein (RYBP) binds to RING1A and RING1B to assist PRC localisation to their targets (Garcia et al. 1999). Interestingly, and possibly due to its interaction with RING1, *Rybp* deletion in mESCs induces the expression of MERVL and two-cell stage specific genes (Hisada et al. 2012) (Figure 1.12). Consistent with the remodeling of H3K27me3 during the MZT, dysregulation of this histone

modification in mESC after interfering with the Chromodomain Helicase DNA binding protein-5 (CHD5) also leads to MERVL derepression (M. Hayashi et al. 2015) (Figure 1.12). Although it remains to be determined how many of these MERVL regulators act directly in the activation of a 2C-like transcriptional program, these studies highlight that the mechanism of epigenetic ZGA regulation can be faithfully recapitulated in an *in vitro* system such as mESCs, facilitating functional studies that are, otherwise, inaccessible in the early embryo. In fact, recently, a systematic screen of chromatin regulators whose siRNA knock-down increases the 2C-like subpopulation identified the non-canonical PRC1 complex PRC1.6 and the E1A-binding protein p400 (EP400) / histone acetyltransferase Tip60 (TIP60) complex as repressors of the emergence of 2C-like cells in mESC cultures (Rodriguez-Terrones et al. 2018) (Figure 1.12).

Interestingly, non-coding RNAs regulate the MERVL-driven transcriptional network not only in early embryos (see section 1.1.6.5) but also in mESCs. The microRNA miR-34a represses MERVL and associated ZGA transcript expression, partially via repression of the GATA-binding factor 2 (GATA2) (Choi et al. 2017) (Figure 1.12). As a result, mESCs deficient of miR-34a show an increased cell fate potential, with extraembryonic contribution in mESC-derived teratomas, embryoid-bodies and chimeric embryos (Choi et al. 2017). Notably, not only GATA2 (Choi et al. 2017), but also GATA3 (Eckersley-Maslin et al. 2019) are inductors of the 2C-like subpopulation (Figure 1.12), consistent with a role for GATA factors in cell fate potency (C. Zhang et al. 2007).

In contrast to all these negative regulators of 2C-like cells and MERVL expression, very few positive inductors have been identified to date. One example is the T-box transcription factor-3 (TBX3) (Dan et al. 2013) (Figure 1.12). TBX3 ectopic expression decreases the levels of DNMT3B and increases TET2, which results in reduced binding of DNMT3B to subtelomeres and subsequently leads to reduced DNA methylation and derepression of genes at subtelomeres, including ZSCAN4 (Dan et al. 2013).

Most notably, the transcription factor DUX was identified as a potent regulator of 2C-like cells and ZGA in embryos (De Iaco et al. 2017; Hendrickson et al. 2017; Whiddon et al. 2017) (see section 1.1.6.4) (Figures 1.9 and 1.12). However, recent *in vivo* evidence shows that lack of

maternal and/or zygotic DUX/DUX4 does not heavily compromise development in mouse or human (Iaco et al. 2019; Z. Chen & Yi Zhang 2019; Vuoristo et al. 2019) and suggests redundancy in ZGA regulation. Consequently, recent studies have aimed to identify maternal regulators of DUX, some of them in a systematic screening manner. This is the case for one of the studies that identified DPPA2 and DPPA4 as critical regulators of ZGA-like transcription in 2C-like cells (Eckersley-Maslin et al. 2019) (Figure 1.12). This candidate-based screen identified another ten chromatin and epigenetic factors that increase the 2C-like subpopulation, as measured by a MERVL reporter and by bulk RNA-sequencing analysis of ZGA transcripts, including Dux (Eckersley-Maslin et al. 2019). These included Interferon Regulator Factor-1 (IRF1), EYA transcriptional coactivator and phosphatase-1 (EYA1), SP110 nuclear body protein (SP110), GATA3, Tumour Protein-63 (TRP63), Bromo Adjacent Homology Domain containing-1 (BAHD1), Ubiquitin Specific Peptidase-3 (USP3), TOX high mobility group box family member-3 (TOX3), Histone Deacetylase-9 (HDAC9) and ZSCAN4C (Eckersley-Maslin et al. 2019) (Figure 1.12).

*Dppa2* and/or *Dppa4* KO mESC completely abrogate the 2C-like subpopulation and expression of MERVL, ZSCAN4 and other ZGA genes and associated retrotransposons, such as LINE-1 (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Y.-L. Yan et al. 2019). As expanded in chapter 5 of this dissertation, DPPA2 and DPPA4 require DUX to initiate 2C-like transcription, suggesting they are upstream regulators (Eckersley-Maslin et al. 2019; De Iaco et al. 2019). Interestingly, it was recently proposed that DPPA2 expression is regulated by sumoylation, specifically, by the SUMO2 E3 ligase PIAS4, whose knock-down induces the 2C-like subpopulation and ZGA-like expression (Figure 1.12), indicating that the sumoylated form of DPPA2 is unable to trigger transcription of *Dux* and ZGA (Y.-L. Yan et al. 2019).

Other recently identified regulators of DUX expression and 2C-like cells are the Growth Arrest and DNA Damage-inducible proteins-45 (GADD45) (Schule et al. 2019) (Figure 1.12). Triple KO of GADD45A/B/G results in downregulation of 2C-like genes, which can be rescued by DUX overexpression (Schule et al. 2019).

Further to the candidate-based screen performed by Eckersley-Maslin et al. (Eckersley-Maslin et al. 2019), recent genome-wide Clustered Regularly Interspaced Short Palindromic Repeat

(CRISPR)-based screens (see sections 1.4 and 1.5) have further attempted to uncover the regulation of ZGA *in vitro* in 2C-like cells. In fact, an independent identification of DPPA2 came through a genome-wide CRISPR-inhibition screen of PIAS4 knock-down mESCs using a ZSCAN4C reporter (Y.-L. Yan et al. 2019). A similar screen design, but using a CRISPR KO approach on a DUX-induced 2C-like subpopulation labeled with a MERVL reporter, was performed by Fu and colleagues (X. Fu et al. 2019), leading to the identification of both positive and negative regulators of the 2C-like subpopulation, including MYC Proto-Oncogene, BHLH Transcription Factor (MYC) and DNMT1 as repressors of *Dux* transcription and 2C-like expression (X. Fu et al. 2019) (Figure 1.12).

Despite all these efforts to understand ZGA regulation, a comprehensive understanding is still missing. This is partly due to the reliance on single or double MERVL and ZSCAN4 reporters to identify ZGA-like cells *in vitro*. The recent advances on sequencing technologies, especially at the single-cell level, promise to serve as a powerful tool to identify ZGA signatures in mESCs that go beyond MERVL and ZSCAN4. The combination of these technologies with genetic perturbation strategies in mESCs constitutes an unprecedented approach for a systematic study of the regulation of ZGA.

## 1.3 Single-cell RNA-sequencing technologies

In recent years, the development of single-cell RNA-sequencing (scRNA-seq) has not only enabled the dissection of variable and heterogenous cell populations, but also the transcriptional profiling of lowly-abundant and rare cell types, which are of special relevance in early development (reviewed in Serena Liu & Trapnell 2016; Hadjantonakis & Arias 2016; Svensson et al. 2018). Initially, plate-based technologies, including methods such as CEL-seq (Hashimshony et al. 2012; Hashimshony et al. 2016), Quartz-seq (Sasagawa et al. 2013), Smart-seq2 (Picelli et al. 2013; Picelli et al. 2014), MARS-seq (Jaitin et al. 2014), MATQ-seq (Sheng et al. 2017), and others (Tang et al. 2009; Tang et al. 2010; Islam et al. 2011; Ramsköld et al. 2012; Islam et al. 2012; X. Fan et al. 2015; Chapman et al. 2015; Vitak et al. 2017; Cao et al. 2017) were developed. These methods are characterised by high transcript detection level and coverage, low bias, high accuracy, but low throughput (reviewed in X.-T. Huang et al.

2018; Birnbaum 2018). However, recent development of combinatorial indexing strategies has allowed increased throughput of plate-based technologies (Vitak et al. 2017; Cao et al. 2017; Rosenberg et al. 2018). Current international efforts aim to profile transcriptionally every cell type of multi-cellular developing and adult organisms, out of which, of special interest, are mouse and human (Pijuan-Sala et al. 2019; Cao et al. 2019; Behjati et al. 2018; Vento-Tormo et al. 2018; Schaum et al. 2018; Regev et al. 2017). To this end, high-throughput scRNA-seq strategies were subsequently developed using droplet-based technologies, which allow profiling the transcriptome of thousands of individual cells at the same time by individually encapsulating them in nanoliter-sized aqueous droplets, each associated with a barcode that gives cell identify (Streets et al. 2014; Macosko et al. 2015; Klein et al. 2015; H. C. Fan et al. 2015; Bose et al. 2015; G. X. Y. Zheng et al. 2017; Zilionis et al. 2017; Gierahn et al. 2017).

In this dissertation, I use droplet-based scRNA-seq technologies to probe candidate maternal regulators of ZGA. Therefore, in the following sections I revise, firstly, how scRNA-seq has been used to profile the transcriptome of early embryos and 2C-like cells and secondly, how droplet-based technologies can be used to profile the transcriptome of thousands of cells in a cost- and time-effective manner.

#### 1.3.1.1 Single-cell RNA-sequencing profiling of early embryos

The development of single-cell technologies was a major advancement for developmental biology due to the limited number of cells available for analysis and the cell-to-cell heterogeneity that drives crucial decisions in the embryo (reviewed in Hadjantonakis & Arias 2016). Most scRNA-seq methods rely on the use of oligo(d)T adaptors to capture mRNAs via their polyadenylated tail at the 3'end and are therefore not suitable to study non-coding transcription. Using these methods, the coding transcriptome of both mouse and human pre-implantation development has been profiled at single-cell resolution (Tang et al. 2011; Xue et al. 2013; L. Yan et al. 2013; Deng et al. 2014; Biase et al. 2014; Blakeley et al. 2015; Boroviak et al. 2018), giving unprecedented insights into the MZT, ZGA and first lineage decisions. Polyadenylated lncRNAs have also been recently profiled using these datasets both in mouse and human pre-implantation development (K. Zhang et al. 2014; Hamazaki et al. 2015; Lv et al. 2015; Qiu et al. 2016). Interestingly, plate-base polyA-capture scRNA-seq technologies

have enable mapping the emergence of the transient and cycling sub-population of 2C-like cells in mESC cultures (Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018; X. Fu et al. 2019) (see section 1.2).

However, recent technological advances have made possible the capture of the whole transcriptome from a single-cell (Sheng et al. 2017; T. Hayashi et al. 2018; Verboom et al. 2019) and, specifically, the small non-coding transcriptome (Faridani et al. 2016). Similar low-input methods have been applied to a pool of embryos, although not to single cells, to profile early embryogenesis (S.-J. Park et al. 2015; Abe et al. 2015), including circular RNAs (X. Fan et al. 2015). Profiling the MZT at single-cell resolution without a bias for polyadenylated transcripts will be of particular importance given the post-transcriptional regulation of maternal transcripts via cytoplasmic polyadenylation (Paynton & Bachvarova 1994; Racki & Richter 2006).

Given the relatively low number of cells in pre-implantation mouse and human embryos, plate-based technologies are preferentially used to map these stages and achieve high transcript coverage without the need to profile thousands of cells. However, deciphering the wide variety of cell types in later post-implantation stages requires high-throughput methods, such as droplet-based technologies, which have been recently used to generate maps of early mouse gastrulation and organogenesis (Pijuan-Sala et al. 2019; Cao et al. 2019) and of the human maternal-fetal interface in first trimester pregnancies (Vento-Tormo et al. 2018).

Unfortunately, these scRNA-seq methods lose important spatial information. To overcome this limitation, recent efforts have aimed to adapt traditional *in situ* hybridisation-based techniques to measure both the abundance and the localisation of hundreds of transcripts in single cells (Lubeck et al. 2014; K. H. Chen et al. 2015; Moffitt et al. 2016; Moffitt & Zhuang 2016; Salmen et al. 2018; Eng et al. 2019; Rodriques et al. 2019), and, excitingly, these have begun to be applied to early embryos (Peng et al. 2019).

The epigenome is now also accessible at single-cell resolution (reviewed in Kelsey et al. 2017), providing much of our current understanding of the molecular events of the MZT (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018). Most notably, multi-omics single-cell sequencing, including simultaneous profiling of DNA methylation and transcription

(Angermueller et al. 2016), DNA methylation and chromatin accessibility (F. Guo et al. 2017; Lin Li et al. 2018) and DNA methylation, chromatin accessibility and transcription (Clark et al. 2018; Argelaguet et al. 2019) has just begun to provide unprecedented insights into the coordination and interdependencies of the transcriptome and the epigenome in early mouse and human embryos.

#### 1.3.1.2 Droplet-based technologies for high-throughput scRNA-seq

In this dissertation, I use scRNA-seq to profile the transcriptome of hundreds of thousands of mESCs where the expression of hundreds of maternal epigenetic and transcription factors was induced using CRISPR technology. As described in chapters 3 to 5 of this dissertation, multiple single cells had to be sequenced per CRISPR perturbation to achieve statistical power, and therefore, a high-throughput droplet-based method was required.

One of the first attempts to prepare cDNA from single-cells using microfluidic technology came from Streets and colleagues (Streets et al. 2014) and consisted on an adaptation of the first plate-based scRNA-seq protocol (Tang et al. 2009) using a microfluidic platform for single-cell manipulation, which improved detection sensitivity and measurement precision (Streets et al. 2014). Consistently, implementation of the widely-used Smart-seq2 protocol (Picelli et al. 2013; Picelli et al. 2014) to the commercially available microfluidic platform Fluidigm C1 improved mRNA capture efficiency (Islam et al. 2014). After these initial proof of concept studies, two protocols, Drop-seq and inDrop, showed effective and specific mRNA capture using beads coated with cell-specific barcodes containing oligo(d)T primers and unique molecular identifiers (UMIs) dispensed into microfluidics droplets (Macosko et al. 2015; Klein et al. 2015; Zilionis et al. 2017). These protocols can routinely sequence tens of thousands of cells in a cost-effective manner, but the number of mRNAs captured per cell is considerably reduced compared to plate-based methods (Ziegenhain et al. 2017). Additionally, the cell capture rates in Drop-seq and inDrop is only 10%, making these protocols unsuitable to profile rare cell populations or samples where the material is limited (Macosko et al. 2015; Klein et al. 2015; Zilionis et al. 2017). Two years later, a commercially available solution, 10X Genomics Chromium, was made available to profile the 3' transcriptome at single-cell resolution (G. X. Y. Zheng et al. 2017). Although the cost per cell is slightly higher than in

Drop-seq or InDrop, 10X Genomics single-cell sequencing is highly convenient for several reasons: 1) it uses a commercial droplet device, avoiding the need to set up manually a sophisticated microfluidic platform; 2) the capture rate is between 50% to 65%, depending on the cell type; 3) the process of encapsulating cells into droplets is incredibly fast, with up to 80,000 cells processed in less than 10 minutes, making it time-effective and reducing the time between sample preparation and droplet formation, thereby preserving cell quality; and 4) it is highly scalable as up to 8 different samples with up to 10,000 cells each can be run in parallel. Consequently, 10X Genomics scRNA-seq is the current method of choice for many laboratories world-wide (reviewed in Birnbaum 2018). In parallel, Seq-Well technology was released, consisting on a portable platform where capture beads and single cells are captured and sealed in an array of subnanoliter wells using a semipermeable membrane (Gierahn et al. 2017). Although Seq-Well was shown to preserve high cell quality, commercially available options have taken over given the simpler experimental set-up. Despite 10X Genomics, which has recently released a 5' capture method (https://www.10xgenomics.com), all these described droplet-based approaches rely on transcript capture with oligo(d)T-coated beads, making them unsuitable to study non-coding or nascent transcription.

In this dissertation, I am interested in detecting the well-defined ZGA-like signature in mESC upon inducing perturbations with CRISPR technology and therefore, I can trade sensitivity in transcript detection with throughput and use a method that allows profiling hundreds of thousands of cells in a cost- and time-effective manner. For these reasons, the method of choice was 10X Genomics 3' scRNA-seq. In this protocol, 10X barcoded gel beads are loaded into a microfluidics system (called Chromium controller) that encapsulates cells with barcodes in oil droplets or gel beads-in-emulsion (GEMs) (Figure 1.13). Each bead is coated with many oligos, each comprising a partial Illumina read 1 (R1) adaptor used for library preparation, a 16-nucleotide 10X cell barcode unique in each bead, a 10-nucleotide UMI and an oligo(d)T to capture the polyadenylated tail of mRNAs (Figure 1.13). After passing through the Chromium controller, a suspension of single cells encapsulated in oil droplets with beads is collected and after cell lysis, reverse transcription and oil removal, a single tube containing cell- and transcript-barcoded cDNA is obtained (Figure 1.13). Lastly, barcoded cDNA is processed via a standard scRNA-seq library preparation protocol (see Materials and Methods) and

downstream analysis allows transcriptional profiling of thousands of genes in individual cells (Figure 1.13).

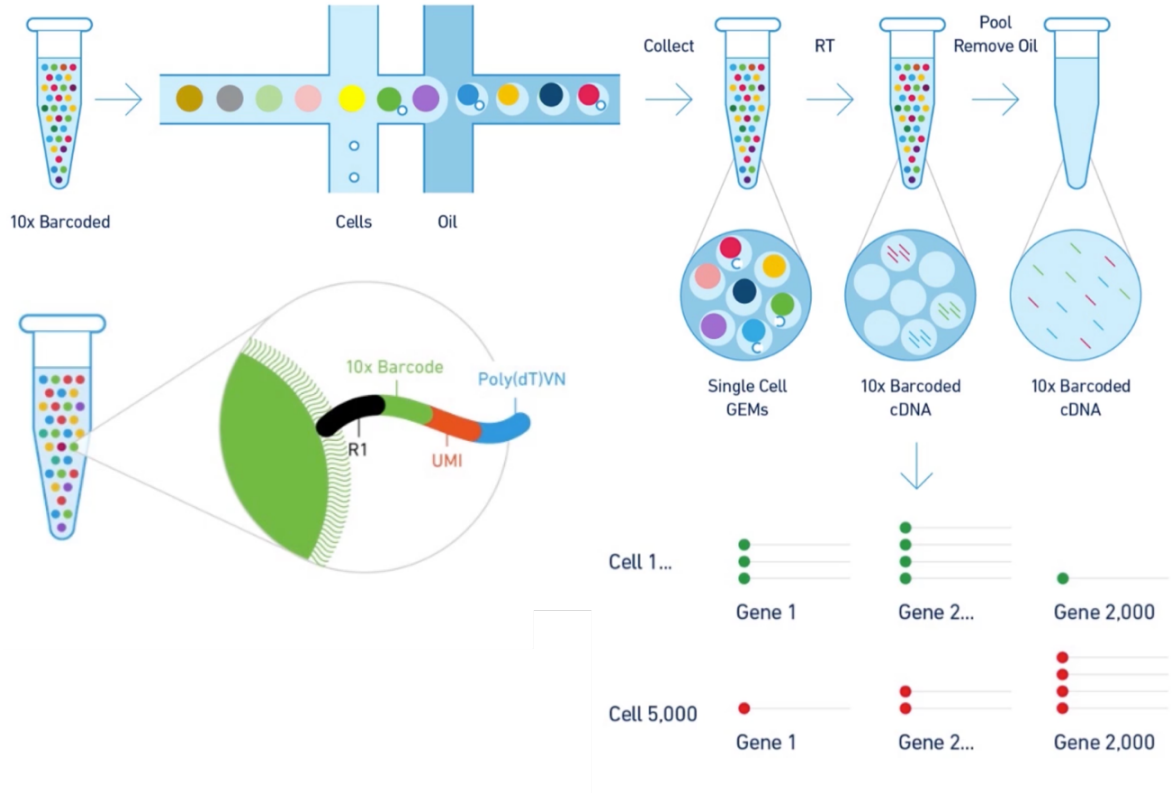



**Figure 1.13– Work-flow of 10X Genomics scRNA-seq droplet formation and transcriptional analysis**

Up to 750,000 different 10X barcoded gel beads are loaded into a microfluidics system (or Chromium controller) together with cells, enzymes (or reagents) and oil. Each cell is encapsulated with a 10X barcoded bead in gel bead-in-emulsions (GEMs). Each bead is coated with many oligos, each comprising a 22-nucleotide partial Illumina Read 1 (R1) adaptor used for library preparation, a 16-nucleotide 10X cell barcode unique in each bead, a 10-nucleotide unique molecular identifier (UMI) and an oligo(d)T (poly(dT)VN) to capture the polyadenylated tail of mRNAs. After passing through the Chromium controller, a suspension of single cells encapsulated in oil droplets with beads is collected and, after cell lysis, reverse transcription (RT) and oil removal, a single tube containing cell- and transcript-barcoded cDNA is obtained. Lastly, barcoded cDNA is processed via a standard scRNA-seq library preparation protocol and downstream analysis allows transcriptional profiling of thousands of genes in individual cells. Image adapted from (https://www.10xgenomics.com).

## 1.4 CRISPR as a tool for genome editing

In the early 1990s, Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) arrays were discovered by three independent groups in bacteria and archea (Ishino et al. 1987; Hermans et al. 1991; Mojica et al. 1993) and later proposed to serve as an adaptive immune system against foreign nucleic acids, including viruses and plasmids (Mojica et al. 2005; Barrangou et al. 2007) (Figure 1.14). In 2002, CRISPR-associated (*Cas*) genes were discovered as RNA-guided DNA endonucleases associated with the clustered repeats (Jansen et al. 2002). In contrast to type I CRISPR systems, type-II systems require only one Cas protein called Cas9 (Makarova et al. 2015; Heler et al. 2015; Wei et al. 2015). In naturally occurring type II-CRISPR systems in bacteria and archea, the invading nucleic acid sequence is incorporated into the host bacterial genome between an array of CRISPR sequences, and these CRISPR repeat arrays are transcribed and processed into CRISPR RNAs (crRNAs) (Brouns et al. 2008) (Figure 1.14). Between the CRISPR arrays, short stretches of conserved DNA, called protospacer adjacent motifs (PAMs), are crucial for the defence mechanism (Bolotin et al. 2005; Mojica et al. 2009) (Figure 1.14). In *Streptococcus pyogenes*, PAM motifs have the form of "NGG", where "N" denotes any nucleotide and "G" denotes guanine (Makarova et al. 2015). Trans-activating crRNA molecules (tracrRNAs) are required to generate mature crRNAs and to bridge them with Cas9 proteins (Deltcheva et al. 2011; Jinek et al. 2012). When a new foreign nucleic acid enters the cell, the crRNA/Cas9 complexes scans the foreign genome for the presence of PAM and, if the adjacent region to it matches the spacer of the crRNA, Cas9 catalyses a double-strand break that halts the infection (Garneau et al. 2010; Jinek et al. 2012; Gasiunas et al. 2012) (Figure 1.14).

It was not until 2012 that these seminal discoveries led to the proposal that CRISPR/Cas9 systems could be used as a tool for genome editing. This was based on biochemical *in vitro* studies that showed that Cas9 could be reprogrammed to target any genomic site of interest that was adjacent to a PAM and induce a double strand break by using a crRNA sequence complementary to the target site of interest (Jinek et al. 2012; Gasiunas et al. 2012). Furthermore, one of these reports showed that the crRNA and the tracrRNA molecules could

be fused together to create a single, synthetic 20-nucleotide guide RNA (sgRNA) that guides and targets Cas9 to the site of interest (Gasiunas et al. 2012).

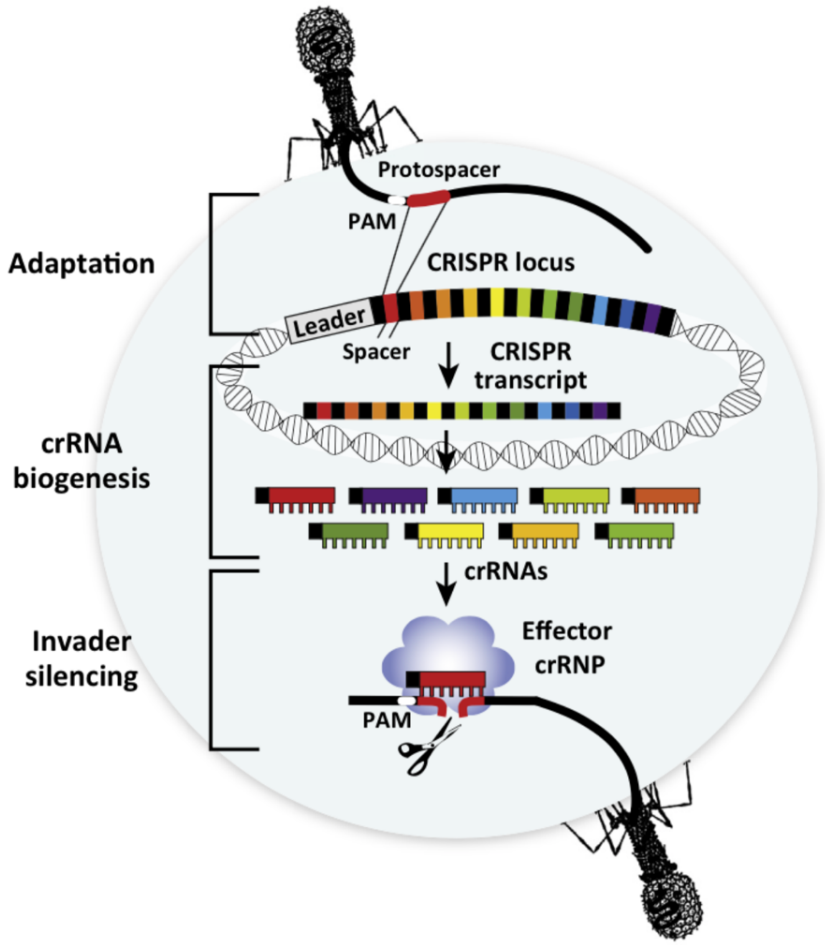



**Figure 1.14– CRISPR/Cas adaptive immune mechanism in bacteria and archea**

Bacteria and archea use CRISPR/Cas as an adaptive immune system against foreign nucleic acids, such as invading viruses. First, in a process called adaptation, the invader DNA is acquired and incorporated into the host CRISPR locus near a sequence called leader. The CRISPR locus contains copies of a short direct repeat sequence (black) in between the invader-derived sequences (multiple colours). The invading sequence selected for integration into the CRISPR locus are called protospacer (red) and are located adjacent to protospacer adjacent motifs (PAMs, white). Transcription of the CRISPR locus generates multiple crRNAs in a process called crRNA biogenesis, where mature crRNAs retain the repeat sequence (black) adjacent to the invader-derived guide sequence (multiple colours). When a new invading sequence comes in contact with the hosts, crRNAs, associated to Cas proteins, recognise the foreign nucleic acid via base-pairing and the Cas protein cleaves in the region of hybridisation, a process called invader silencing. In this process, PAMs are critical for recognition and silencing of the invading sequence. Image from Terns & Terns 2014.

A year later, two independent studies reported that CRISPR/Cas9 from *S. pyogenes* and *S. thermophilus* could be expressed in mouse and human cells to cleave a target site in a programmable manner and that it could induce homology-directed repair (Cong et al. 2013; Mali, Yang, et al. 2013). Therefore, since it only requires a 20-nucleotide sgRNA complementary to a region of interest containing a PAM and expression of Cas9, CRISPR is an incredibly useful tool for genome editing. Consequently, since 2013, CRISPR has been installed in the scientific community as a powerful tool for genome editing used by multiple laboratories world-wide with a plethora of applications, from the generation of transgenic animals, to lineage molecular recording and even the first clinical applications (reviewed in Pickar-Oliver & Gersbach 2019).

### 1.4.1 Endogenous transcriptional activation using CRISPR: CRISPRa

Multiple CRISPR applications rely on the inactivation of the catalytic activity of Cas9 to render a dead Cas9 (dCas9), by introducing two mutations (D10A and H841A in *S. pyogenes*) in the nuclease domains HNH and RuvC1 (Jinek et al. 2012; Gasiunas et al. 2012; Qi et al. 2013; Konermann et al. 2013; Mali, Aach, et al. 2013). The use of a dCas9 repurposes CRISPR to a sequence-specific RNA-guided platform that can be used to activate or inhibit transcriptional activity (CRISPR-activation or CRISPRa and CRISPR-inhibition or CRISPRi, respectively) (Mali, Aach, et al. 2013; Gilbert et al. 2013; A. W. Cheng et al. 2013; Perez-Pinera et al. 2013; Maeder et al. 2013; Larson et al. 2013; Gilbert et al. 2014; Tanenbaum et al. 2014; Konermann et al. 2015; Chakraborty et al. 2014; Chavez et al. 2015; Zalatan et al. 2015; Chavez et al. 2016; Rajagopal et al. 2016), alter the epigenetic state of a gene or genomic element by DNA methylation or histone modifications (Hilton et al. 2015; Kearns et al. 2015; Polstein et al. 2015; Thakore et al. 2015; Vojta et al. 2016; Choudhury et al. 2016; D. Y. Kwon et al. 2017), edit specific nucleotides (Thakore et al. 2015; Hess et al. 2016; Y. Ma et al. 2016) or image genomic regions in living cells (Tanenbaum et al. 2014; reviewed in B. Chen et al. 2016), amongst other applications, simply by fusing dCas9 to different functional domains and designing sgRNAs to target the genomic region of interest (reviewed in Sander & Joung 2014; Adli 2018; Doench 2018) (Figure 1.15).

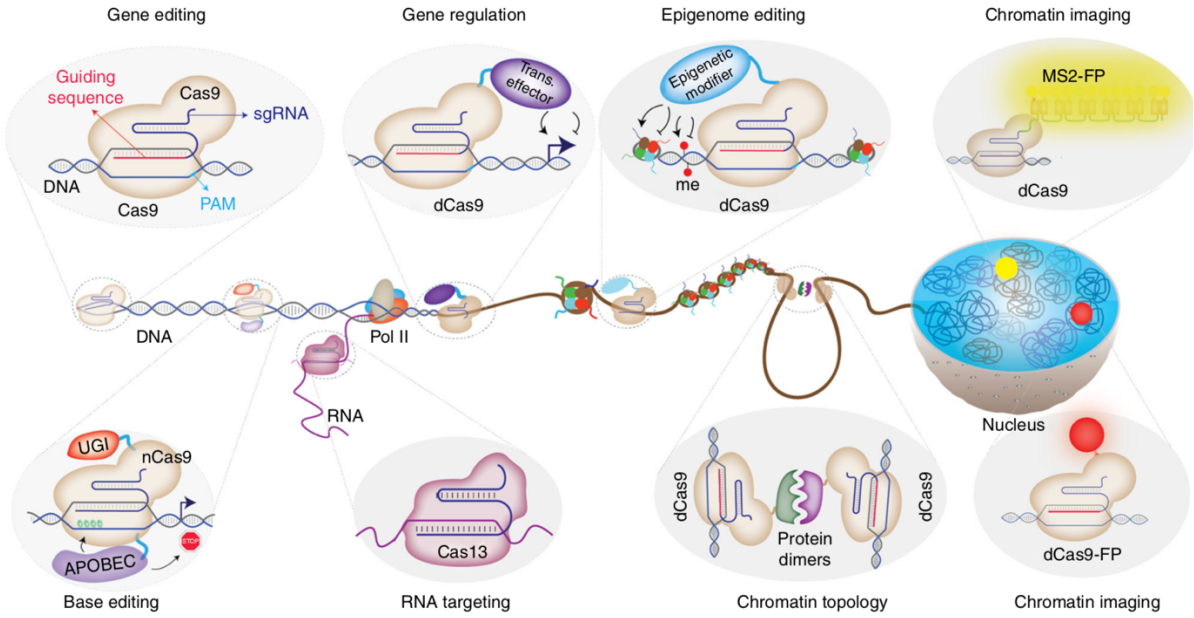



**Figure 1.15– CRISPR applications**

Cas9, in its wild-type form, can be used for gene editing by generating double DNA breaks (top left panel). A catalytically-inactive or dead Cas9 (dCas9) can fused to different domains for multiple purposes: to regulate gene expression, either by activating or repressing it, when fused to different trans-effectors (second top panel from left); to edit epigenetic marks, such as DNA methylation (me), when fused to epigenetic modifiers (third top panel from left); to image chromatin when fused either directly or via MS2 loops to fluorescent proteins (FP) (top and bottom right panels); or to edit chromatin topology when fused to proteins that dimerise (third bottom panel from left). A nickase Cas9 (nCas9) can be used to edit DNA bases in the absence of double DNA breaks when fused to "apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like" (APOBEC) and "uracil DNA-glycolyase inhibitor" (UGI) (bottom left panel). Lastly, Cas13 can be used for RNA-targeting strategies (second bottom panel from left). Pol II: polymerase II. Image from Adli 2018

In this dissertation, I use CRISPRa to induce the endogenous transcription of candidate maternal ZGA regulators in mESCs and assess the transcriptional consequences. CRISPRa can be achieved by fusing dCas9 to different trans-activation domains and designing sgRNAs against the region upstream of the gene TSS (or gene promoter) (Chavez et al. 2016). Recruiting trans-activation domains to gene promoters via dCas9 leads to the recruitment of polymerases that activate transcription (Gilbert et al. 2013). In eukaryotic cells, the first generation of CRISPRa methods, used to induce endogenous gene activation, consisted on fusing dCas9 to the transcription activation domain tetrameric VP16 (also called VP64) or to the nuclear factor NF-kappa-β p65 subunit (p65) (Gilbert et al. 2013). VP16 is a transcription factor from the herpex simplex virus involved in the expression of viral genes by interacting

with numerous host factors to recruit RNA polymerase II and binding histone modifiers that allow transcriptional activation (Tumbar et al. 1999). In CRISPRa applications, VP16 domain is commonly found fused to dCas9 in four copies for enhanced activity, and this is referred to as VP64. The activation domain p65, from the NF-kappa-β signalling pathway, is involved in innate and adaptive immune responses and interacts with members of the basal transcription complex, such as TATA-binding protein and Transcription factor II β, making it a potent transcription inducer (van Essen et al. 2009). VP64 fusion to dCas9 is more efficient than p65 fusion (Gilbert et al. 2013) and, therefore, the first generation of CRISPRa methods used dCas9-VP64 fusion (Figure 1.16). Although this first generation of activators has been used in a number of studies, showing that it can activate both silent and already active genes, the fold change achieved compared to basal gene expression is very modest (Mali, Aach, et al. 2013; Gilbert et al. 2013; A. W. Cheng et al. 2013; Perez-Pinera et al. 2013; Maeder et al. 2013). The observation that using multiple sgRNAs to recruit additional activators enhanced target gene activation (A. W. Cheng et al. 2013; Maeder et al. 2013) led to the development of the second generation of activators (reviewed in La Russa & Qi 2015).

One of the second generation CRISPRa methods is SunTag, developed by two independent groups (Gilbert et al. 2014; Tanenbaum et al. 2014). In SunTag CRISPRa system, dCas9 is fused to a tandem array of peptides to recruit many copies of the VP64 effector, which in turn is fused to a single-chain variable fragment (scFv) that recognises the peptide array (Figure 1.16). This method can lead to an increase of up to 50-fold in the protein levels of the targeted gene (Tanenbaum et al. 2014).

Another potent CRISPRa strategy consists on employing multiple different activators fused to dCas9 in tandem, namely VP64, p65 and Replication and Transcription activator (RTa) (Chavez et al. 2015). RTa is an immediate-early murine gammaherpesviruses gene product found to activate viral lytic genes by binding to RTa-responsive elements in the promoter regions of target genes (Buisson et al. 1989). This method, known as VPR, has been successfully applied in human, mouse, *Drosophila melanogaster* and *Saccharomyces cerevisiae* cells to activate target genes and upregulate their mRNAs from 5- to 300-fold using multiple different sgRNAs per target (Chavez et al. 2015).

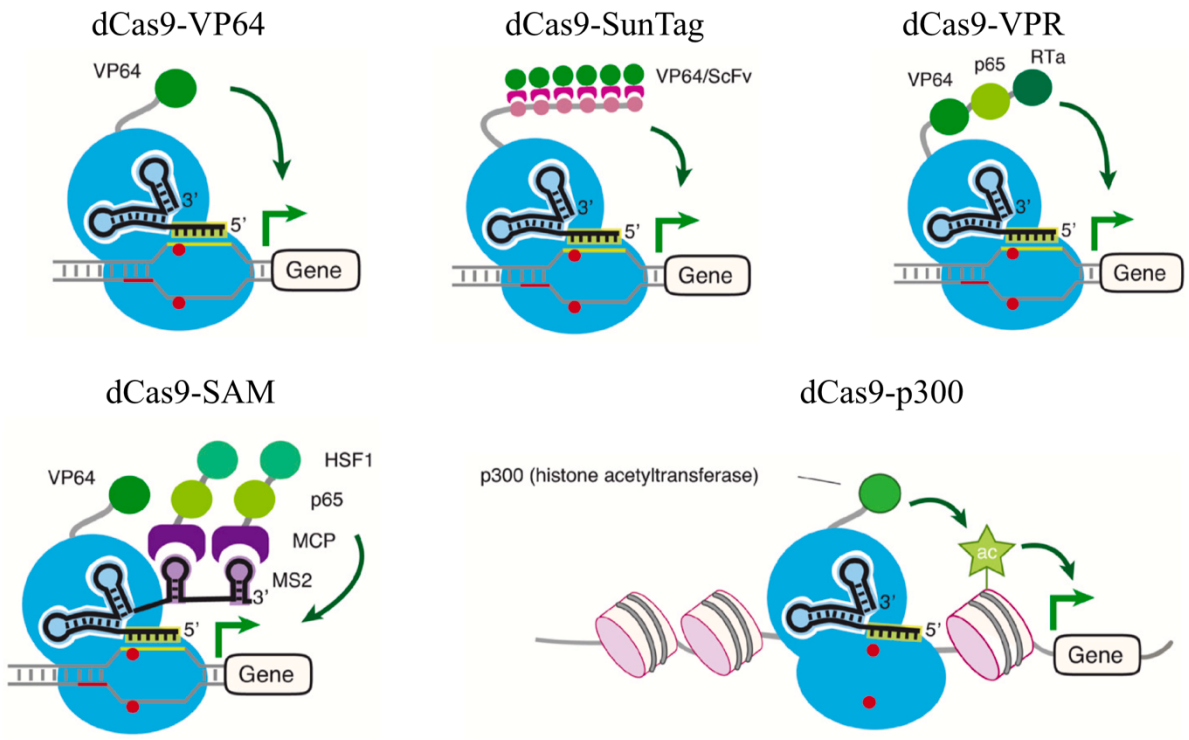



**Figure 1.16– CRISPRa strategies**

Transcriptional activation of endogenous loci (CRISPRa) can be achieved by tethering dead Cas9 (dCas9) with different trans-activation domains. Different methods have been developed: fusion of dCas9 to VP64 (top left); SunTag, in which multiple copies of VP64 are recruited by interaction between single-chain variable fragments (scFv, fused to VP64) with a peptide array (fused to dCas9) (top middle); VPR, in which dCas9 is fused to VP64, p65 and RTa (top right); SAM, in which dCas9 is fused to VP64 and MCP-p65-HSF1 fusions are recruited via MS2 loops in the scaffold sequence of the sgRNA (bottom left); or epigenome editing methods such as dCas9 fusion to the histone acetyltransferase p300 (bottom right). VP64: tetrameric VP16 transcription activation domain; RTa: replication and transcription activator; SAM: synergistic activator mediator; HSF1: heat shock factor-1 protein; MCP: MS2 coat protein; ac: acetylation. Image from Lo & Qi 2017.

Interestingly, a third CRISPRa approach called synergistic activator mediator (SAM) can achieve similar or even higher levels of target gene upregulation than SunTag and VPR using a single sgRNA (Konermann et al. 2015). This strategy also uses three different trans-activators, however, the synergistic effect is achieved by using two different fusion proteins: 1) dCas9-VP64 and 2) p65 fused together with heat shock factor-1 protein (HSF1) to the MS2 RNA binding protein (also known as MCP or MS2 coat protein) (Konermann et al. 2015) (Figure 1.16). HSF1 is a transcription factor and the major mediator of the heat shock response, by inducing transcription of heat shock proteins in response to environmental stress (reviewed in Dayalan Naidu & Dinkova-Kostova 2017). To allow recruitment of p65 and HSF1 to the

dCas9-VP64/sgRNA complex at the promoter of the gene of interest, the scaffold sequence of the sgRNA is modified to include two MS2 loops (Figure 1.16). MS2 RNA binding proteins bind to MS2 loops as dimers, therefore, further to VP64, four additional copies of activation modules are recruited to the targeted promoter, leading to mRNA fold changes from two- to multiple-thousand-fold, depending on basal gene expression levels of the targeted gene (Konermann et al. 2015).

Lastly, fusion of the histone acetyltransferase p300 to dCas9 can be used to induce endogenous gene transcription via deposition of H3K27ac by designing sgRNAs against enhancers, instead of promoters, creating an epigenome editing platform that can be used to upregulate gene expression via indirect mechanisms (Hilton et al. 2015) (Figure 1.16). However, this strategy requires detailed knowledge of the regulation of cell-type specific enhancers and, therefore, it is less adaptable than previously described strategies.

A recent study comparing SunTag, VPR and SAM CRISPRa systems across several human, mouse and fly cell lines showed that SAM is the most potent and consistent in inducing target gene activation, while maintaining high specificity (Chavez et al. 2016). Consequently, in this dissertation, I decided to employ SAM as a robust method for CRISPRa screening.

CRISPRa has a number of advantages over traditional cDNA overexpression techniques. Firstly, the levels of gene expression achieved are more physiologically-relevant by CRISPRa (Chavez et al. 2015; Sanson et al. 2018; J. Yang et al. 2019). Secondly, it does not require cloning of the cDNA of interest, making it amenable for transcriptional upregulation of any gene, regardless of its size. Thirdly, it can be multiplexed and highly scalable, allowing simultaneous activation of multiple genes in a single pooled experiment, making it very suitable for screening purposes (Konermann et al. 2015; Joung et al. 2017; Horlbeck et al. 2016; Sanson et al. 2018).

### 1.4.2 Pooled genetic screens using CRISPR

Genetic screening refers to the assessment of multiple gene functions in a systematic way. Pooled screening means the simultaneous testing of thousands of individual perturbations in a single batch, and contrasts to arrayed screening in which each perturbation is tested

individually in a single well (Figure 1.17). Pooled screening has multiple advantages over arrayed screening: the set-up is simpler as it does not require automation or specialised equipment for liquid handling, they are high-throughput allowing comprehensive profiling of genome-wide libraries, and they are cost-effective. However, pooled screening also has some disadvantages, such as the requirement of high number of cells, the biosafety considerations for viral preparation and the dependency on next-generation sequencing (NGS) (reviewed in J. Yao & Dai 2018).

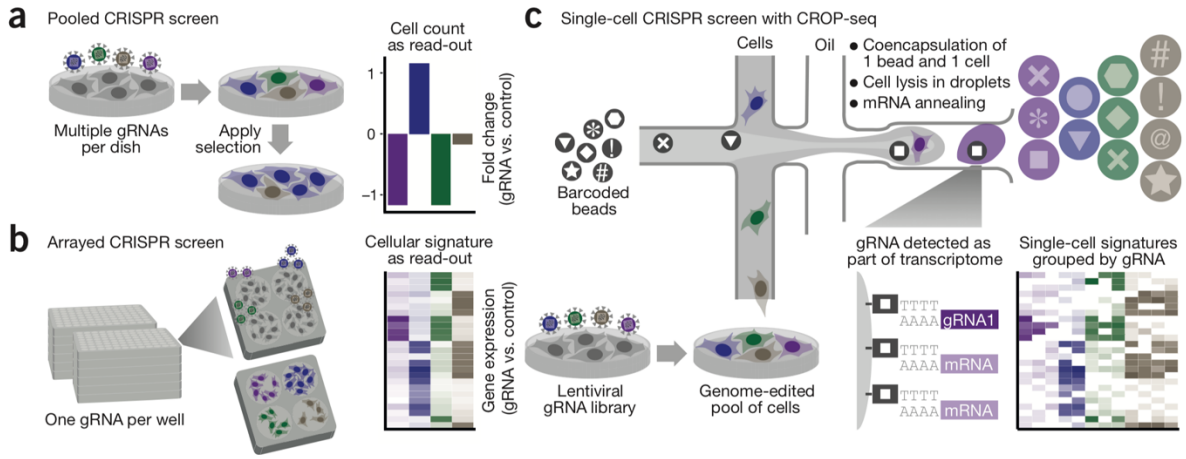



**Figure 1.17– Pooled and arrayed CRISPR screening strategies**

Pooled CRISPR screens can be done by delivering a lentiviral sgRNA library into cells and, after applying some sort of selection, changes in sgRNA abundance amongst bulk populations of cells are assayed by next-generation sequencing analysis on the extracted DNA (a). In contrast, arrayed CRISPR screens support complex read-outs, such as transcriptional cellular signatures, but require assaying each perturbation separately, making them low-throughput and labour-intensive (b). Pooled CRISPR screening with single-cell read-out, using methods such as CROP-seq, combines the advantages of a) and b), allowing pooled lentiviral delivery and complex analysis of each perturbation by single-cell RNA-sequencing, where the transcriptome of each cell is profiled individually along with the expressed sgRNA (c). Image from Datlinger et al. 2017.

Before the discovery of CRISPR as a tool for genome editing, mammalian genetic screening was routinely carried out with RNAi technology, where phenotypes were interpreted after loss-of-function perturbations. However, the rate of off-target effects mediated by RNAi are relatively high and the gene suppression achieved is often insufficient to allow observation of KO phenotypes (reviewed in Mohr & Perrimon 2012). Over the last years, the simplicity and specificity of CRISPR for genome editing has led to an explosion in the application of CRISPR for genetic pooled screening. In short, a library of sgRNAs (genome-wide or targeted to a set

of genes or genomic features of interests) is delivered to cells and, after some sort of selection, the perturbations that cause the phenotype of interest are analysed by NGS of the genomic DNA to quantify the abundance of each perturbation (Gasperini et al. 2016; reviewed in Doench 2018) (Figure 1.17). In these multiplexed CRISPR screenings, replication-inactivated lentiviruses are often the delivery method of choice for sgRNA libraries (Figure 1.17), since they are stably integrated into the host genome, can infect both dividing and non-dividing cells and are relatively easy to produce and manipulate to control the number of integrations per cell (reviewed in Lino et al. 2018).

CRISPR screens can be classified based on the type of perturbations introduced. CRISPR screening was first performed in mammalian cells by introducing Cas9-mediated loss-of function mutations via double-strand DNA breaks in coding regions of the genes of interest (Shalem et al. 2014; T. Wang et al. 2014; Koike-Yusa et al. 2014) (Figure 1.15). These are called CRISPR KO screens and are widely used both in genome-wide and targeted screens in an ample variety of systems (reviewed in Doench 2018). Loss-of-function screening can also be performed using CRISPRi, which is carried out by directing dCas9 alone to the region around the TSS of the genes of interest (Qi et al. 2013; Gilbert et al. 2013) or fused to the transcriptional repressor KRAB (Gilbert et al. 2013; Gilbert et al. 2014; Horlbeck et al. 2016; Radzisheuskaya et al. 2016; Sanson et al. 2018) (Figure 1.15). Interestingly, gain-of-function screens can also be performed using second generation CRISPRa systems (Figure 1.16). One of the first genome-wide CRISPRa screens was performed using the SunTag system (Gilbert et al. 2014). By using a tiling sgRNA library, the authors found that optimal CRISPRa efficiency was achieved with sgRNAs targeted 50 to 400 bp upstream of the TSS regardless of basal gene expression levels (Gilbert et al. 2014). Consequently, they designed a genome-wide library using this targeting rule and screened for genes that affect cell growth after activation in the human K562 cell line (Gilbert et al. 2014). Interestingly, the same screen design using CRISPRi rendered different results: CRISPRi hits were limited to expressed genes whereas CRISPRa hits included genes across a broad range of endogenous expression levels (Gilbert et al. 2014), therefore highlighting the power of CRISPRa for screening gene function regardless of the basal expression state of genes. The study that first reported the development of CRISPRa SAM technology (Konermann et al. 2015) showed optimal activation efficiency with

sgRNAs targeted 200 bp upstream of the gene's TSS as well as successful multiplex gene activation using this system. A genome-scale screen using a SAM lentiviral library containing 3 sgRNAs per gene, targeted 200 bp upstream of the TSS, was successful in identifying gain-of-function perturbations that conferred resistance to a small molecule inhibitor in a melanoma cell line (Konermann et al. 2015). More recently, further optimised genome-wide and targeted CRISPRa libraries have been constructed (Horlbeck et al. 2016; Sanson et al. 2018) based on two observations: 1) CRISPR/Cas9 is influenced by nucleosome occupancy (Horlbeck, Witkowsky, et al. 2016; Horlbeck et al. 2016), and 2) functional TSS annotation with the FANTOM database (Forrest et al. 2014), which identifies TSSs using Cap Analysis of Gene Expression (CAGE), is more accurate for functional sgRNA design than RefSeq or Ensembl annotations, in both CRISPRi and CRISPRa applications (Radzisheuskaya et al. 2016; Horlbeck et al. 2016; Sanson et al. 2018).

Based on their read-out, pooled CRISPR screens can be divided into viability screens, screens based on flow-cytometry read-outs and screens based on single-cell read-outs (reviewed in Doench 2018). Viability screens rely on affecting cell fitness, so that perturbations that affect it are depleted or enriched by the end of the screen. This type of screens is widely used in cancer biology to identify dependencies of tumour cells on a particular mutation, and are often combined with drug screening to unravel mutations that confer resistance. CRISPR screening can be combined with flow cytometry to separate cells of interest based on fluorescent reporters incorporated in the cells or stained with fluorescent antibodies. However, one of the critical premises of pooled CRISPR screening is maintaining adequate representation of the library at all steps and, therefore, one of the main drawbacks of CRISPR screening based on flow-cytometry read-outs is the extensive hours of sort time required, especially for large-scale or genome-wide screens. In these two types of screens, the populations of cells separated based on survival or expression of fluorescent markers are subsequently analysed by NGS on the extracted DNA. This way, sgRNAs constitute a unique barcode that can be used to unravel enriched or depleted perturbations compared to sgRNA non-targeting controls and associated with the phenotype of interest (Figure 1.17). The third type of CRISPR screens, those based on single-cell read-outs, are significantly different from the other two and were recently made possible thanks to the combination of CRISPR technology with single-cell sequencing

(reviewed in Doench 2018; J. Yao & Dai 2018) (Figure 1.17). In this case, analysis of the whole transcriptome or epigenetic state of each single cell is used to understand the phenotype caused by each perturbation. While CRISPR KO and CRISPRi screens based on single-cell read-outs have been previously performed, in this dissertation, I developed for the first time a combination of pooled CRISPRa screening with scRNA-seq read-out to probe maternal epigenetic and transcription factors as regulators of ZGA. Therefore, single-cell read-out-based screens are reviewed in detail in the following section 1.5.

## 1.5 CRISPR screening at single-cell resolution

Until recently, complex CRISPR screening molecular read-outs, such as transcriptional or epigenetic profiling, were limited to arrayed screens (Tan & S. E. Martin 2016) (Figure 1.17), making them very labour-intensive and low-throughput in the number of perturbations that could be tested in one experiment. Combining pooled CRISPR screening with complex read-outs is not straight forward, since it implies being able to identify the perturbation introduced in every cell that renders the observed transcriptional and/or epigenetic changes (Figure 1.17). Even though single-cell sequencing presented itself as the obvious method to analyse these read-outs, an important limitation became apparent: sgRNAs are often transcribed from a U6 promoter which is bound by polymerase III (Pol III), leading to a non- 3' polyadenylated transcript (H. Ma et al. 2014). As reviewed earlier in section 1.3, most scRNA-seq methods rely on the capture of 3' polyadenylated transcripts. Since 2016, a number of methods have overcome this limitation and shown that it is possible to combine CRISPR screening with single-cell-based read-outs, namely CROP-seq (Datlinger et al. 2017), Perturb-seq (Dixit et al. 2016; Adamson et al. 2016; Replogle et al. 2018; Rubin et al. 2019) and CRISP-seq (Jaitin et al. 2016). These methods not only allow performing high-throughput pooled CRISPR screenings with complex interrogation of gene function but also permit disentangling cell-to-cell heterogeneities (Figure 1.17).

CROP-seq implements a smart cloning strategy that allows simultaneous analysis of the whole transcriptome and detection of the sgRNA expressed in each cell (Datlinger et al. 2017) (Figure 1.17). The strategy consists on incorporating the sgRNA transcript expressed in each cell within

a polymerase II (Pol II) transcript that can be capture in scRNA-seq libraries (Figure 1.18). For that, the authors took advantage of the replication mechanism of lentiviruses and incorporated the U6 Pol III promoter and sgRNA cassette within the 3'LTR, at the same position of the 400 bp deletion that makes lentivirus self-inactivating, and just before the 3' LTR polyadenylation signal (Figure 1.18). Upstream of the 3' LTR, contained within the lentiviral integration site, a Pol II promoter drives expression of an antibiotic-resistance cassette used for selection of transduced cells (Figure 1.18). Upon lentiviral reverse transcription and integration into the host genome, the U6-sgRNA cassette is copied to the 5' LTR (Figure 1.18). Functional sgRNAs are then expressed in the host cell from the U6 promoter but also, the U6-sgRNA cassette is contained within the Pol II polyadenylated transcript, making it detectable by scRNA-seq approaches (Figure 1.18). Using this strategy, sgRNA expression can be assigned to every single-cell and downstream transcriptional changes induced by each perturbation can be determined (Figure 1.17). Interestingly, CROP-seq has been used to probe not only genes but also enhancer elements, as demonstrated recently (Gasperini et al. 2019).

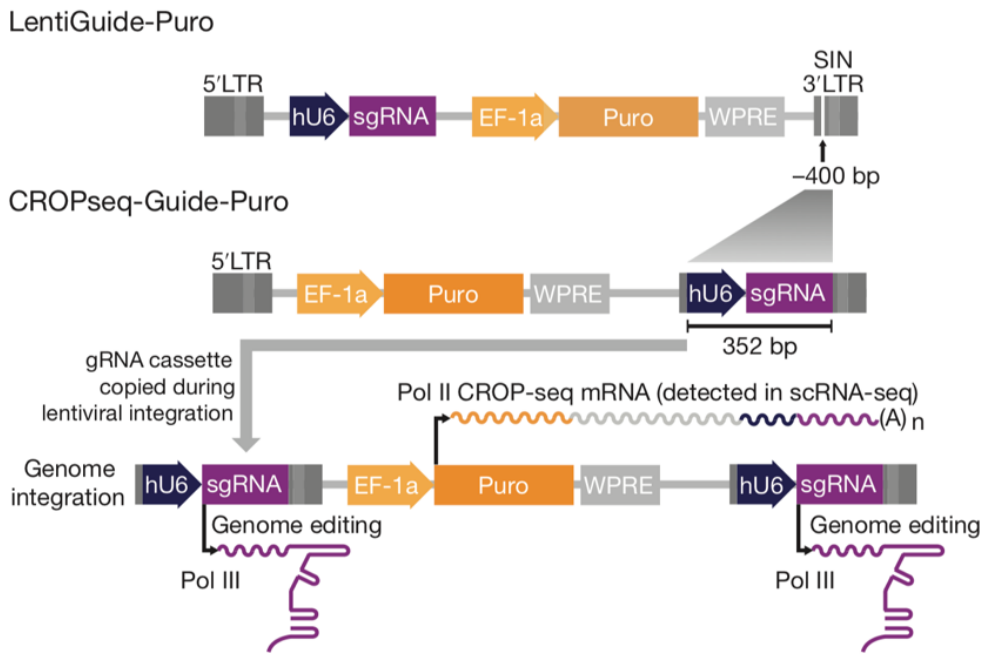



**Figure 1.18– Schematic representation of CROP-seq strategy**

The CROPseq-Guide-Puro was derived from the LentiGuide-Puro by moving the hU6-sgRNA cassette within the 3' LTR. After lentiviral integration and transcription from the host genome, the sgRNA is contained within a polyadenylated Pol II transcript detectable by single-cell RNA-sequencing (scRNA-seq), while a functional sgRNA Pol III transcript is also produced. Image from Datlinger et al. 2017.

In contrast, Perturb-seq (Dixit et al. 2016; Adamson et al. 2016) and CRISP-seq (Jaitin et al. 2016) methods do not detect the sgRNA sequence itself by scRNA-seq but rather an associated polyadenylated barcode previously incorporated in the sgRNA libraries. Using a very similar strategy, Mosaic-seq was used to systematically study the function of enhancer elements (S. Xie et al. 2017). Although these methods have been shown to be highly scalable, a critical drawback has been pointed out in screens performed during the last year: uncoupling of the sgRNA-barcode pairs can occur during lentiviral packaging of pooled libraries, leading to inaccurate sgRNA-cell assignments after scRNA-seq (Hill et al. 2018; S. Xie et al. 2018; Adamson et al. 2018; Feldman et al. 2018). Even though alternatives have been proposed to reduce this uncoupling, including arrayed packaging of lentiviral libraries (Adamson et al. 2018; Feldman et al. 2018), they come with a substantial reduction in throughput and scalability, making CROP-seq a more suitable strategy for high-throughput pooled CRISPR screening (Hill et al. 2018). However, the disruption of the 3' LTR used in CROP-seq implies that only one sgRNA can be cloned per lentiviral vector (Datlinger et al. 2017), making this option unsuitable for multiple sgRNA delivery in a single cell, unless a high multiplicity-of-infection (MOI) is used for random combinatorial integrations of sgRNAs into cells. Nevertheless, the experimental design presented in this dissertation only required one perturbation introduced per cell and, therefore, I opted to adapt CROP-seq for pooled CRISPR screening.

Last year, an alternative strategy was proposed to combine the best features of CROP-seq and Perturb-seq/CRISP-seq methods to capture directly the non-polyadenylated sgRNA sequence in 10X Genomics scRNA-seq libraries (Replogle et al. 2018). For that, this method, called direct-capture Perturb-seq, requires the incorporation of custom sequences in the stem loop or at the 3' end of the sgRNA scaffold sequence in the sgRNA libraries. Importantly, these modifications in the sgRNA sequence do not compromise sgRNA activity (Replogle et al. 2018). At the time of scRNA-seq, 10X beads are not only coated with oligo(d)T adaptors but also with specialised barcodes that recognize the modified sgRNA scaffold and consequently, capture both mRNAs and sgRNA transcripts (Replogle et al. 2018). Since this method is commercially available, it is likely to become the method of choice for pooled CRISPR scRNA-seq screening.

Most scRNA-seq-based CRISPR screens performed to date have employed loss-of function perturbations through CRISPR KO (Datlinger et al. 2017; Jaitin et al. 2016; Dixit et al. 2016) or CRISPRi (Adamson et al. 2016; S. Xie et al. 2017; Replogle et al. 2018; Genga et al. 2019; Gasperini et al. 2019) to screen different biological functions of genes (Datlinger et al. 2017; Jaitin et al. 2016; Dixit et al. 2016; Adamson et al. 2016; Replogle et al. 2018; Genga et al. 2019) and enhancers (S. Xie et al. 2017; Gasperini et al. 2019). Therefore, these approaches can only be used to interrogate genes that are already expressed or enhancers that are already active in the cellular system under study. Recently, two months before the deposition of this dissertation, CRISPRa has been combined with scRNA-seq read-out using Perturb-seq to generate genetic interaction maps of cell fitness in human cell lines (Norman et al. 2019). However, the combination of CRISPRa with single-cell read-outs as a screening tool is presented for the first time in this dissertation.

Excitingly, not only the transcriptome but also epigenetic features, such as chromatin accessibility measured by ATAC-seq, can also now be combined with pooled CRISPR screening at single-cell resolution (Rubin et al. 2019). Even though the scalability of these methods is still limited and genome-scale screens are not feasible due to the single-cell sequencing costs, these methods promise great advents in the understanding of mammalian gene function.

# Chapter 2 Materials and Methods

## 2.1 Cloning

### 2.1.1 Vectors

The following vectors were used in this dissertation:

Table 2.1 – Vectors used, providers and short description

| Vector name | Provider | Short description |
|---|---|---|
| pMD2.G | Addgene (12259) | It expresses the human immunodeficiency virus (HIV) VSV-G envelope was used for lentiviral packaging together with psPAX2 and a lentiviral vector expressing the sequence of interest. |
| psPAX2 | Addgene (12260) | It expresses the HIV genes *gag*, *pol*, *rev*, and *tat* and was used for lentiviral packaging together with pMD2.G and a lentiviral vector expressing the sequence of interest. |
| Lenti dCas9-VP64_Blast | Addgene (61425) | Third generation lentiviral vector that expresses the fusion dCas9-VP64 and was used to generate SAM mESCs (section 3.2.1). |
| Lenti MS2-p65-HSF1_Hygro | Addgene (61426) | Third generation lentiviral vector that expresses the fusion MS2-p65-HSF1 and was used to generate SAM mESCs (section 3.2.1). |

| | | |
|---|---|---|
| Lenti sgRNA(MS2)_ puro | Addgene (73795) | Third generation lentiviral backbone for sgRNA cloning containing a scaffold sequence with MS2 loops at the tetraloop and stemloop-2, and an elongation factor-1 alpha (EF1α)-puromycin resistance marker. It also contains BsmBI sites for insertion of a sgRNA spacer sequence. It was used in experiments described in sections 3.2.2 and 3.2.3. |
| Lenti CROP-sgRNA-MS2 | Custom vector built by VectorBuilder | See section 2.1.2. Full sequence provided in Appendix A. |
| pDONR221 | Thermo Fisher Scientific (12536017) | Gateway plasmid used to clone in the PCR products of the cDNA sequences used for experiments described in sections 5.2.3, 5.2.4 and Appendix J. |
| pIG400 | In-house built | Gateway pDEST vector containing a CAG promoter and an in-frame C-terminal eGFP coding sequence linked to a blasticidin resistance marker by an internal ribosome entry site (IRES) sequence. It was used to express GFP fusion proteins after transferring the relevant cDNA sequence from pDONR221 by Gateway cloning, in the experiments described in sections 5.2.3, 5.2.4 and Appendix J. Full sequence provided in Appendix B. |

### 2.1.2 Synthesis of CROP-sgRNA-MS2 lentiviral construct

The CROP-sgRNA-MS2 lentiviral vector was synthesised by VectorBuilder using the design of the CROPseq-Guide-Puro (Addgene, 86708) (Datlinger et al. 2017) with the following modifications: 1) the sgRNA scaffold sequence contains two MS2 loops, as in the sgRNA(MS2)_puro backbone (Addgene, 73795), that allow recruitment of MS2-p65-HSF1 in

SAM22 mESCs; and 2) a fluorescent mCherry marker was included downstream of the EF1α promoter and linked through T2A to a puromycin resistance cassette, allowing assessment of the MOI by flow cytometry and antibiotic selection of the cells (Figure 3.27). This vector's sequence is provided in Appendix A.

### 2.1.3 Cloning of individual sgRNAs

For individual sgRNA cloning, two oligos were synthesised per sgRNA (Sigma Aldrich), one containing the spacer or target sequence with a "CACCG" flank at the 5' end and the other one synthesised as the reverse complementary sequence to the target sequence and flanked by "AAAC" at the 5' end and by a "C" at the 3' end. Each oligo pair was annealed using T4 Polynucleotide Kinase (PNK) enzyme (NEB, M0201S) and then cloned into the sgRNA(MS2)_puro backbone (Addgene 73795) or into the in-house built CROP-sgRNA-MS2 backbone (Figure 3.27) by a Golden Gate reaction using BsmBI enzyme (Thermo Fisher Scientific, ER0451) and T7 ligase (NEB, M0318S). The product from the Golden Gate reaction was transformed into Stbl3 competent cells (Thermo Fisher Scientific, C737303). Between 2 to 3 colonies were picked per sgRNA and verified by Sanger sequencing. All sgRNA target sequences used in this dissertation were design in Joung et al. 2017 and sequences are provided in Appendix C.

### 2.1.4 Cloning of the 475 sgRNA pooled library

For cloning the 475 sgRNA pooled library into the CROP-sgRNA-MS2 backbone, first, oligos containing the sgRNA target sequence with a 5' end 26 base-pair (bp) flanking region complementary to the U6 promoter (TATCTTGTGGAAAGGACGAAACACCG) and a 3' end 35 bp flanking region complementary to the sgRNA scaffold sequence (GTTTAAGAGCTAGGCCAACATGAGGATCACCCATG) were synthesised by Twist Bioscience. This oligo library was then amplified and cloned into the CROP-sgRNA-MS2 vector using Gibson assembly, as previously described in Joung et al. 2017, by VectorBuilder. Library coverage was estimated to be >11,000 folds by colony count of diluted transformations. 150 bp paired-end sequencing was performed on Illumina HiSeq4000 to analyse sgRNA representation in the library. The target sgRNA sequences synthesised and cloned into the

CROP-sgRNA-MS2 lentiviral backbone and the number of sequencing reads for each sgRNA in the final pooled plasmid library are provided in Appendix C.

### 2.1.5 Cloning of cDNA-eGFP constructs

Sequence-verified cDNA sequences lacking stop codons (Appendix D) were polymerase chain reaction (PCR) amplified from plasmids purchased from Thermo Fisher Scientific using forward primers containing an AttB1 sequence (GGGGACAAGTTTGTACAAAAAAGCAG GCTTCACC) followed by the 22 5'end bases of the cDNA of interest, and reverse primers containing an AttB2 sequence (GGGGACCACTTTGTACAAGAAAGCTGGGTC) followed by the 24 complementary bases to the 3' end of the cDNA sequence (Table 2.2). The PCR was performed with Phusion High-Fidelity DNA Polymerase (NEB, M0530S) or LongAmp Tap DNA Polymerase (NEB, M0323), following manufacturer's instructions. The PCR product was then cloned into pDONR221 vector (Thermo Fisher Scientific, 12536017) using a Gateway BP clonase II enzyme (Thermo Fisher Scientific, 11789020), following manufacturer's instructions. A Gateway LR II clonase reaction (Thermo Fisher Scientific, 11791100) was then used to transfer the cDNA sequences into an in-house built pDEST vector (pIG400) containing a CAG promoter and an in-frame C-terminal eGFP coding sequence linked to a blasticidin resistance marker by an IRES sequence (full plasmid sequence provided in Appendix B). Expression plasmids were sequence verified by Sanger Sequencing prior to use.

Table 2.2 – Primer sequences for PCR amplification of cDNA sequences

| **Gene** | **Forward primer (5' flanking GGGGACAAGTTTGTACAAAAAAGCAGGCTTCACC)** | **Reverse primer (5' flanking GGGGACCACTTTGTACAAGAAAGCTGGGTC)** |
|---|---|---|
| *Dppa2* | ATGTCATACTTCGGCCTGGAGA | CTACGGCGGCATATTTGGGGGTAG |
| *Smarca5* | ATGTCGTCCGCGGTGGAGCCTC | TAGTTTCAGCTTCTTTTTTCTTCC |
| *Patz1* | ATGGAGCGGGTCAACGACGCTT | CTTCCCTTCAGGCCCCATGGGCTG |
| *Carhsp1* | ATGTCATCTGAGCCTCCTCCAC | GTTGCTGATGACGTGCCCAGACCA |

## 2.2 Cell culture

### 2.2.1 Culture of mouse embryonic stem cells (mESCs)

All mouse embryonic stem cells (mESCs) used in this dissertation were grown under serum/LIF conditions: DMEM (Gibco, 11995-040), 15% fetal bovine serum (FBS), 1 U/mL penicillin - 1 mg/mL streptomycin (Gibco, 15140-122), 0.1 mM nonessential amino acids (Gibco, 11140-050), 4 mM GlutaMAX (Gibco, 35050-061), 50 μM β-mercaptoethanol (Gibco, 31350-010), and $10^3$ U/mL LIF (Stem Cell Institute, Cambridge). mESCs were cultured on tissue culture plates coated with immortalised mouse embryonic fibroblast (MEFs) or with 0.1% gelatin, and maintained in an incubator at 37℃ with 5% $CO_2$ humidified atmosphere. Media was refreshed every day and the cells passaged every other day with Trypsin EDTA (Thermo Fisher Scientific, 25200056).

### 2.2.2 Culture of HEK293T

HEK293T cells were grown in D10 media: DMEM (Gibco, 11995-040), 10% FBS, 1 U/mL penicillin - 1 mg/mL streptomycin (Gibco, 15140-122). HEK293T cells were cultured on T175 tissue culture flasks or 100 mm tissue culture plates and maintained in an incubator at 37℃ in 5% $CO_2$ humidified atmosphere. Media was refreshed every other day and the cells passaged every three days with Trypsin EDTA (Thermo Fisher Scientific, 25200056).

### 2.2.3 Lentiviral packaging

#### 2.2.3.1 Individual lentiviruses

Individual lentiviruses were packaged for experiments in chapter 3 and validation experiments in chapter 5 (section 5.2.3) and Appendix J. First, 3.5 million HEK293T cells were seeded into 100 mm tissue culture plates 24 hours before transfection. The day after, they were cotransfected with 3.5 μg of pMD2.G (Addgene, 12259), 6.5 μg of psPAX2 (Addgene, 12260) and 10 μg of the lentiviral vector of interest: dCas9-VP64_Blast (Addgene, 61425), MS2-p65-HSF1_Hygro (Addgene, 61426), sgRNA(MS2)_puro (Addgene, 73795) cloned with an

individual sgRNA, or CROP-sgRNA-MS2 cloned with an individual sgRNA. A single-tube reaction mix was prepared for each transfection containing the three lentiviral plasmids and 60 µl of TransIT Reagent (Mirus Bio, 2700) diluted in 1.5 mL of Opti-MEM (Gibco, 31985), which was subsequently added drop-wise into the cells cultured in 8.5 mL of fresh media, following manufacturer's instructions. 48 hours later, 10 mL of viral supernatant were harvested by filtering through a 0.45 µm filter (Sartorius, 16533). For sgRNA-expressing lentivirus from the CROP-sgRNA-MS2 backbone, the 10 mL of viral supernatant were concentrated to 800 µL with LentiX Concentrator (Takara, 631231), following manufacturer's instructions, due to the lower viral titer (Datlinger et al. 2017). Before transduction, supernatants were supplemented with 8 µg/mL polybrene (Millpore, TR-1003-G)

#### 2.2.3.2 Pooled sgRNA lentiviral library

VectorBuilder packaged into lentivirus the 475 sgRNA pooled library cloned into the CROP-sgRNA-MS2 lentiviral vector. Briefly, the plasmid library was co-transfected with their proprietary envelop plasmid encoding VSV-G and packaging plasmids encoding *Gag*, *Pol* and *Rev* into HEK293T packaging cells. After a short incubation period, the supernatant was collected, followed by removal of cell debris by centrifugation, filtration and polyethylene glycol (PEG) concentration of the viral particles. To measure lentiviral titer, HEK293T cells were transduced with lentivirus diluted from the stock. Then, a quantitative reverse transcription PCR (qPCR)-based approach was used to quantify the average number of integration events of the proviral genome per host genome to estimate titer in the original viral stock. From the titer provided, I then performed the titration on SAM22 mESCs as described in the result section 4.2.2.

### 2.2.4 Lentiviral transduction and generation of cell lines

#### 2.2.4.1 Generation of SAM7 and SAM22 mESCs

SAM7 and SAM22 mESCs were generated by lentiviral transduction of lenti dCas9-VP64_Blast (Addgene, 61425) and lenti MS2-p65-HSF1_Hygro (Addgene, 61426) into E14 mESCs, followed by antibiotic selection and manual subcloning. Briefly, 1.5 million E14

mESCs were seeded into 100 mm tissue culture plates and, the day after, they were supplemented with 5 mL of a 1:1 lentiviral mixture of dCas9-VP64 and MS2-p65-HSF1 (5 mL out of the 20 mL combined viral mixture, see section 2.2.3.1). 24 hours later, cells were washed with phosphate-buffered saline (PBS) and subsequently cultured in mESC media supplemented with 20 μg/mL blasticidin (InvivoGen, ant-bl-05) and 200 μg/mL hygromycin (InvivoGen, ant-hg-1). After 10 days of antibiotic selection, 18 individual colonies were picked manually and expanded to 24-well plates before DNA isolation (see section 2.3).

#### 2.2.4.2 Generation of sgRNA-expressing cell lines from the sgRNA (MS2)_puro backbone

Lentiviral transduction of sgRNAs expressed from the sgRNA(MS2)_puro backbone was performed on SAM22 mESCs previously cultured for at least 5 days in mESC culture media containing 20 μg/mL blasticidin (InvivoGen, ant-bl-05) and 200 μg/mL hygromycin (InvivoGen, ant-hg-1) to ensure expression of dCas9-VP64 and MS2-p65-HSF1. After collection of the corresponding lentiviral supernatant in 10 mL (see section 2.2.3.1), 1.5 million SAM22 mESCs were supplemented with 1 mL of lentivirus diluted in 9 mL of mESC media, without blasticidin or hygromycin, and seeded in 100 mm tissue culture plates. 24 hours after transduction, the cells were re-infected with 1 mL of lentivirus diluted in 9 mL of regular mESC culture media. 24 hours later, cells were washed with PBS and media was replaced to mESC media containing 1 μg/mL puromycin. After 6-15 days of antibiotic selection and culture, cells were frozen down or harvested for RNA isolation, western-blotting, immunofluorescence analysis or 10X Genomics scRNA-seq.

#### 2.2.4.3 Generation of sgRNA-expressing cell lines from the CROP-sgRNA-MS2 backbone

Lentiviral transduction of sgRNAs expressed from the CROP-sgRNA-MS2 backbone was performed on SAM22 mESCs previously cultured for at least 5 days in mESC culture media containing 20 μg/mL blasticidin (InvivoGen, ant-bl-05) and 200 μg/mL hygromycin (InvivoGen, ant-hg-1) to ensure expression of dCas9-VP64 and MS2-p65-HSF1. After

collection and concentration of the corresponding lentiviral supernatant to 800 µL (see section 2.2.3.1), 120,000 SAM22 mESCs were supplemented with all the volume of the concentrated lentivirus and seeded in one well of a 24-well plate. 24 hours later, cells were washed with PBS and media was replaced to regular mESC culture media. Two days later, media was replaced to 1 µg/mL puromycin-containing mESC media and cells were cultured for an extra 7 days before being frozen down or harvested for RNA isolation, immunofluorescence analysis or 10X Genomics scRNA-seq.

#### 2.2.4.4 Transduction of the pooled sgRNA lentiviral library

Detailed description of the number of SAM22 mESCs and lentiviral volumes used for transduction of the pooled 475 sgRNA lentiviral library is provided in the result section 4.2.2. The transduction was performed by direct supplementation of the lentiviral supernatant into the cells.

### 2.2.5 Other cell lines used

*Dppa2* knock-out (KO) mESCs were previously described in Eckersley-Maslin et al. 2019 and *Smarca5* (*Snf2h*) KO mESCs in Barisic et al. 2019.

## 2.3 DNA isolation and genomic PCR

DNA from clones of E14 mESCs transduced with dCas9-VP64 and MS2-p65-HSF1 (see section 2.2.4.1) was isolated using AllPrep DNA/RNA mini Kit (Qiagen, 80004). Clones were screened for dCas9-VP64 and MS2-p65-HSF1 genomic insertion by PCR using DreamTaq DNA Polymerase (Thermo Fisher Scientific, P0701) and primers provided in Table 2.3. The resulting PCR products were run in a 1% agarose gel with HyperLadder IV (Biolane, BIO-33029) (Figure 3.3).

Table 2.3 – Primer sequences for genomic PCR of dCas9-VP64 and MS2-p65-HSF1

| Insertion | Forward primer | Reverse primer |
|---|---|---|
| dCas9-VP64 | CATCGAGCAGATCAGCGAGT | CGATCCGTGTCTCGTACAGG |
| MS2-p65-HSF1 | AAGCCTGAACTCACCGCTAC | TTGGGAATCCCCGAACATGG |

## 2.4 RNA isolation, cDNA synthesis and quantitative reverse transcription PCR (qPCR)

RNA was isolated using RNeasy Mini kit (Qiagen, 74104), RNAasy Micro kit (Qiagen, 74004) or AllPrep DNA/RNA mini Kit (Qiagen, 80004) and treated with DNaseI (Thermo Fisher Scientific, EN0521) following manufacturer's instructions. cDNA was synthesised from 0.5-2 µg of DNAaseI-treated RNA using RevertAid First-Strand cDNA Synthesis Kit (Thermo Fisher Scientific, K1622), and diluted 1:10 prior to qPCR. qPCRs were performed in two technical replicates using Brilliant III SYBR master mix (Agilent Technologies, 600882) and a CFX384 Touch Real-Time PCR Detection System machine (BioRad). Relative levels of transcript expression were quantified by the comparative CT method with normalisation to *Gapdh* levels. Primer sequences are provided in Table 2.4.

Table 2.4 – Primer sequences for qPCR

| Transcript | Forward primer | Reverse primer |
|---|---|---|
| sgRNA | GTTTTAGAGCTAGGCCAACATGA (amplify sgRNA scaffold sequence) | TGGCCAAGTTGATAACGGACT (amplify sgRNA scaffold sequence) |
| *Gapdh* | GGTGGTGAAGCAGGCATCT | CGGCATCGAAGGTGGAAGA |
| MERVL | CTTCCATTCACAGCTGCGA | GTGCTTCGCATACAGTGCAG |
| *Zscan4* | AGTCTGACTGATGGTGCT | GGCCTTGTTGCAGATTGCT |

| | (amplify all isoforms) | (amplify all isoforms) |
|---|---|---|
| *Gm4014* | TGTCATGGGCCCAGGTTTAC | AGTACACGGTGCGTTCCTTT |
| *Tcstv3* | GGTTTTCGCCAGAAGTGAGA | GGAGCTGACAGAAGATCCC |
| *Gm8994* | ACGAGATGCCCATGAACCTG | TGGGTCATTTACACGTGCATTA |
| *Zfp352* | CCAGGACCCTGCAATACACA | TACAGGTTGTCTCCTGTGTGC |
| *Gm5039* | ACTTTTGTAAAGAAGGTCCTTTGT | TTCCAAACCCTCAATACAGAAGC |
| *Gm428* and related transcripts | CTTCCCTCTGCATTGGTGGT | TGCTTGATCTTCTGGCCTGG |
| *Gm4340/ Gm2123* | GCCACAGTCTTTTCTTCTGGC | GGACCTCCCCCTCCCTAAAT |
| *Dux* | AGGCCCTGCTATCAACTTTCAAGA | CTGAGACCCCCATTCGCTTG |
| *Pou5f1* | GACCGCCCCAATGCCGTGAA | TGGTCTGGCTGAACACCTTTCCA |
| *Nanog* | CCTCGCCATCACACTGACATGAGT | GCAGGTCTTCAGAGGAAGGGCG |
| *Smarca5* | ACGGTACAAAGCACCCTTTCA | CTGAGGGGAGTTTCGGATGC |
| *Dppa2* | CACTCTTCGGGAGTGGTGTC | TTCTGGCCTCCCGAGATGTA |
| *Eif1a* and related transcripts | GGACACATTTGGTCCTGGAGA | CAAGACACACGCATGCCAAA |
| *Tmem92* | GGTTCTTGGTCATCCTGGCAT | TCCAAGGAGGTGACCCAAATG |
| *Mreg* | CCGCTGGTCAGTGGTAACAA | TCCTATCGTCATCCGCCTCT |
| *Obox6* | ATGTCCGCAAGAGTCCCAAG | CCCTGGAAACCAGCTCTCAC |

## 2.5 Western-blotting

Total cell extracts were prepared in a detergent buffer (10 mM Tris-HCl, pH 7.4, 150 mM NaCl, 10 mM KCl, 0.5% NP-40) containing a protease inhibitor cocktail (Sigma, P2714) and incubated at 4°C for 20 minutes, followed by centrifugation (12,000 × g, 10 minutes). Protein concentration was quantified with Bio-Rad Protein Assay Dye Reagent (Bio-Rad, 5000006). 50 μg of protein were resolved in 4-12% SDS-PAGE gels (Expedon, NBT41212) and blotted on polyvinylidene difluoride (PVDF) membranes. Membranes were blocked in 3% milk in 0.01% Tween20 in PBS and incubated overnight at 4°C with primary antibodies diluted in this same blocking buffer. Primary antibodies and dilutions used were: 1:1000 rabbit anti-ZSCAN4 (Millipore, 2793611), 1:1000 mouse anti-β-ACTIN (Abcam, ab6276), 1:500 mouse anti-DPPA2 (Millipore MAB4356), 1:1000 rabbit anti-OCT4 (Abcam, ab19857), 1:1000 rabbit anti-NANOG (Abcam, ab80891) and 1:5000 mouse anti-HSP90 (Abcam, ab13492). After washing with 0.01% Tween20 in PBS for 45 minutes (3 washes of 15 minutes each), Horseradish peroxidase (HRP)-conjugated secondary antibodies diluted 1:3000 in blocking buffer were incubated for 1 hour and subsequently washed with 0.01% Tween20 in PBS for 1 hour (4 washes of 15 minutes each). Secondary antibodies used were: HRP-conjugated anti-mouse (Santa Cruz Biotechnology, sc-2005) and HRP-conjugated anti-rabbit (Santa Cruz Biotechnology, sc-2004). Detection was carried out with enhanced chemiluminescence (ECL) reaction (GE Healthcare, RPN2209) using standard X-ray films. Band intensities were quantified using ImageJ software.

## 2.6 Immunofluorescence and imaging

### 2.6.1 Immunofluorescence and imaging on mESCs

mESCs were grown on 0.1% gelatin-coated glass coverslips in regular culture conditions and then fixed with 4% paraformaldehyde (PFA) (Polysciences, Inc., 18814) for 20 minutes and permabilised with 0.5% TritonX-100 in PBS for 30 minutes at room temperature. After blocking for 1 hour in 3% bovine serum albumin (BSA) in 0.05% Tween20 in PBS (BS buffer), corresponding primary antibodies diluted in BS were incubated either 1 hour at room

temperature or overnight at 4°C. Primary antibodies and dilutions used were: 1:200 rabbit anti-ZSCAN4 (Millipore, 2793611) and 1:250 rabbit anti-DPPA3 (STELLA) (Abcam, ab19878). After three 10-minutes washes with BS, the secondary antibody donkey anti-rabbit IgG Alexa Fluor (AF)-conjugated 488 (Invitrogen, A21206) diluted 1:1000 in BS was incubated for 30 minutes at room temperature and washed four times, 15 minutes each, with 0.05% Tween20 in PBS. DNA was then counterstained with 5 µg/mL DAPI in PBS and coverslips mounted in ProLong Gold Antidade mounting media (Invitrogen, P36934). Single optical sections were captured with a Zeiss LSM780 microscope (63x oil-immersion objective) and the images pseudo-coloured using Fiji or Adobe Photoshop. For visualisation, images were corrected for brightness and contrast, within the recommendations for scientific data. Fluorescence semi-quantification analysis was performed with Volocity 6.3 (Improvision).

### 2.6.2 Immunofluorescence and imaging on embryos

All mice used in this study were C57BL/6 and were bred and maintained in the Babraham Institute Biological Support Unit. All procedures were covered by a project license to Wolf Reik under the Animal Scientific Procedures Act 1986, and are locally regulated by the Babraham Institute Animal Welfare, Experimentation, and Ethics Committee.

Embryos were collected from females after superovulation and mating to males. Zygotes were collected on the day of plugging and two-cell embryos one day after plugging in M2 media (Sigma-Aldrich, MR-015P-5F) containing hyaluronidase (Sigma, H2126). Subsequently, embryos were washed in M2 droplets to remove the cumulus cells, fixated with 4% PFA (Polysciences, Inc., 18814) for 10 minutes, permeabilised with 0.5% TritonX-100 in PBS for 1 hour and blocked with BS for another hour. Then, primary antibodies were diluted in BS and incubated for 1 hour, followed by 1-hour wash in BS. Next, secondary antibodies diluted in BS were incubated for 45 minutes, followed by 30-minutes to 1-hour wash in 0.05% Tween20 in PBS. DNA was then counterstained with 5 µg/mL DAPI in PBS and embryos mounted in fibrin clots. All incubations were performed at room temperature. Primary antibodies and dilutions used were: rabbit 1:100 anti-SMARCA5 (SNF2H) (Abcam, ab72499) and 1:200 mouse anti-DPPA2 (Millipore, mab4356). Secondary antibodies used were anti-rabbit AF-conjugated 568 and anti-mouse AF-conjugated 488 (Molecular Probes) and diluted 1:1000. Single optical

sections were captured with a Zeiss LSM780 microscope (63x oil-immersion objective) and the images pseudo-coloured using ImageJ2. For visualisation, images were corrected for brightness and contrast, within the recommendations for scientific data. Fluorescence co-localisation analysis was performed with Volocity 6.3 (Improvision) in 10 zygotes and 10 two-cell embryos. Pearson correlation coefficient between SMARCA5 and DPPA2 signals were calculated in the area corresponding to the pronuclei or nuclei. The pronuclei in zygotes and nuclei of each blastomere in two-cell embryos were measured separately, with values comparable within the same embryo.

## 2.7 Flow cytometry and fluorescence-activated cell sorting (FACS)

For the 10X Genomics scRNA-seq experiment presented in section 3.2.3 with E14, and SAM22 mESCs transduced with MERVL and Zscan4 sgRNAs, 8,000 live single-cells from each condition were fluorescence-activated cell sorted (FACS) from a $DAPI^-$ population using a BD FACSAria III Cell Sorter into 1.5 mL collection tubes (Eppendorf, 0030120086) containing 29 μL of 0.1% FBS in PBS. For MOI calculation after transduction of the pooled sgRNA lentiviral library (Figures 4.7 and 4.9), mCherry expression was analysed by flow-cytometry on BD LSR Fortessa from a $DAPI^-$ population. For single-cell NOMe, methylation and transcription-sequencing (scNMT-seq) analysis of *Dppa3* CRISPRa, 44 live SAM22 mESCs stably transduced with the Dppa3 sgRNA 379 (Appendix C) and 44 live SAM22 mESCs stably transduced with the non-targeting sgRNA 462 (Appendix C) were single-cell FACS sorted for $DAPI^-$ on a BD Influx High-Speed Cell Sorter each into a well of a 96-well plate (Eppendorf, 0030129512) containing a GpC methyltransferase mix (see section 2.10). For screen cDNA validations (section 5.2.3 and Appendix J) and mechanistic experiments of *Dppa2* and *Smarca5* (section 5.2.4), 200,000-500,000 $DAPI^-/eGFP^+$ cells were FACS sorted on a BD Influx High-Speed Cell Sorter Sorter into 1.5 mL collection tubes (Eppendorf, 0030120086) and tdTomato fluorescence was also recorded for *Dppa2* WT and KO mESCs (Eckersley-Maslin et al. 2019) (Figure 5.33) and analysed using FlowJo v10.

## 2.8 Bulk RNA-sequencing

### 2.8.1 Preparation of libraries and sequencing

All bulk RNA-sequencing libraries were prepared from 1 µg of DNaseI-treated RNA using Sanger Institute Illumina bespoke pipelines. E14, SAM7 and SAM22 mESC libraries (section 3.2.1) were opposite strand-specific total RNA libraries (ribozero) and sequenced at 100 bp paired-end on the Illumina HiSeq2500 Rapid Run platform. MERVL LTR and *Zscan4* CRISPRa libraries (section 3.2.2) were opposite strand-specific total RNA libraries (ribozero) and sequenced at 50 bp single-end on the Illumina HiSeq2500 Rapid Run platform. *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1* CRISPRa and cDNA overexpression libraries (section 5.2.3), as well as those presented in Appendix J, were opposite strand-specific polyA RNA libraries and sequenced at 50 bp single-end on Illumina HiSeq4000.

### 2.8.2 Data processing and analysis

For processing of all bulk RNA-sequencing data, raw FastQ files were trimmed with Trim Galore (v0.4.1-v0.6.1, default parameters) and mapped to the mouse mm10 genome assembly using TopHat2 (v2.0.12) (D. Kim et al. 2013) or Hisat2 (v2.0.5) (D. Kim et al. 2015), as guided by known splice sites taken from Ensembl v70-v96. Hits were filtered to remove mappings with MAPQ scores <20. For alignments to dCas9-VP64 and MS2-p65-HSF1 exogenous transcripts in E14, SAM7 and SAM22 data (Figure 3.4), two artificial chromosomes were constructed with the sequences between both LTRs for each lentiviral plasmid (Figure 3.2) and reads were mapped as described for mm10 genome assembly.

Data were quantified at mRNA level using the RNA-sequencing quantification pipeline in SeqMonk software (www.bioinformatics.babraham.ac.uk/projects/seqmonk/) with strand-specific quantification using mRNA probes. For dCas9-VP64 and MS2-p65-HSF1 quantification, the reads mapping to the two artificial chromosomes were integrated to the mm10 genome assembly to quantify their expression in relation to the whole transcriptome (Figure 3.4). Differential gene expression analysis were done with different parameters for the

different datasets, as specified in each figure caption, using EdgeR, DESeq2 and/or and intensity difference filters.

For alignments to repetitive regions in the genome, artificial repeat genomes were constructed from repeat annotations downloaded from the UCSC table browser (RepeatMasker, mm10). Sequences of the list of repeat element instances were stitched together separated by 'NNNNN' to create repeat-specific genomes. Trimmed reads from each sample were aligned against all individual repeat genomes using Bowtie2 (v1.0.1-v2.3.3, default parameters) (Langmead & Salzberg 2012). Values reported are cumulative reads mapping to a specific repeat group as percentage of the total read count (Figures 3.10 and 5.23)

## 2.9 10X Genomics single-cell RNA-sequencing (scRNA-seq)

### 2.9.1 Library preparation and sequencing

#### 2.9.1.1 Sample and 10X Genomics V2 library preparation

The sample preparation to obtain a single cell suspension to be loaded into the 10X Genomics Chromium controller was optimised over the course of this dissertation. Briefly, in the first experiment to analyse CRISPRa of MERVL LTR and *Zscan4* (section 3.2.3), cells were washed with PBS, trypsinised as usual, spin down, resuspended in PBS, passed through a 50 μm cell strainer (Thermo Fisher Scientific, NC9667434) and FACS sorted with a DAPI live/dead staining to collect the appropriate number of live cells (see section 2.7). This sample preparation procedure resulted in low cell capture rate (24.5-29.5%) in the 10X libraries (Figure 3.19) with a prominent population of low-quality cells (Figure 3.20). Subsequently, the sample preparation procedure was optimised, resulting in the following protocol which rendered a capture efficiency of 71% (Figure 4.11) and good quality of most of the cells captured (Figure 4.12): cells in culture in 150 mm plates were washed twice with PBS and trypsinised for 2 minutes with 3 mL of Trypsin EDTA (Thermo Fisher Scientific, 25200056); then 5 mL of mESC media were added to the plate and a single cell suspension was obtained by gently pipetting up and down 10-15 times; next, an extra 17 mL of media were added to obtain a final

volume of 25 mL cell suspension which were spin down (300 × g, 3 minutes) and the resulting pellet resuspended in 1 mL of 0.1% FBS in PBS; this cell suspension was then passed twice through 50 μm cell strainers (Thermo Fisher Scientific, NC9667434) and diluted in a total of 5 mL of 0.1% FBS in PBS; cells were then counted in three biological replicates and two technical replicates in a NucleoCounter NC-3000 (Chemometec) using a viability dye; only samples with > 90% viability were taken forward; lastly, the counted cell suspension was spin down (300 × g, 3 minutes) and resuspended in the appropriate volume to obtain a sample concentration of 600 cells / μL, from which 34 μL were loaded in each lane of the 10X Genomics Chromium controller chip to aim to recover 10,000 cells per lane.

In the CRISPRa experiment of MERVL LTR and *Zscan4* (section 3.2.3), a 8,000-cell suspension from each the following samples were loaded each in a lane of the 10X Genomics Chromium Controller chip: E14 mESCs, SAM22 mESCs, SAM22 mESCs transduced with the non-targeting sgRNA control 1, SAM22 mESCs transduced with the MERVL LTR sgRNAs 1 and 3 individually and pooled at the time of sequencing, and SAM22 mESCs transduced with the Zscan4 sgRNAs 1 and 3 individually and pooled at the time of sequencing (see Appendix C for sgRNA sequences). sgRNAs for this experiment were cloned in the sgRNA(MS2)_puro backbone. Droplet formation in the sample of untransduced SAM22 mESCs failed (data not shown) and was not carried forward.

In the CRISPRa experiment from section 3.2.4 using MERVL sgRNA 1, Zscan4 sgRNA 1, Dppa4 sgRNAs 1 and 2, Kdm5b sgRNAs 1 and 2, Dux sgRNAs 1 and 2 and Dppa3 sgRNAs 1 and 2 (see Appendix C for sgRNA sequences), SAM22 mESCs were individually transduced with lentiviruses encoding each of these sgRNAs cloned in the CROP-sgRNA-MS2 backbone and pooled at equal ratios at the time of sequencing. A 10,000-cell suspension of the pool of cells was loaded in one lane of the 10X Genomics Chromium Controller chip.

In the CRISPRa screen experiment of SAM22 mESCs transduced with the pooled 475 sgRNA lentiviral library (chapters 4 and 5), each transduction replicate was loaded across a full 10X Chromium Controller chip (8 lanes), with 20,000 cells per lane. In all 10X Genomics scRNA-seq experiments, libraries were prepared using the 10X V2 Single Cell 3′ Solution kit (10X Genomics) following manufacturer's instructions.

### 2.9.1.2 Amplicon sgRNA libraries

All amplicon sgRNA PCRs were performed on full length cDNA samples obtained during the library preparation protocol of 10X Genomics V2 scRNA-seq libraries, as previously described (Hill et al. 2018). Briefly, 10 ng of full length 10X cDNA were used as starting material and each round of PCR amplification, using with the primers described in Table 2.5 and Figure 3.31, was monitored by KAPA SYBR (Kapa Biosystems, KR0389) to avoid overcycling. The PCR products obtained after each PCR round were cleaned-up with 1X AMPure XP beads (Beckman Coulter, A63881) and eluted in 20 μL. The second and third rounds of PCR were done with 1 μL of a 1:25 dilution of the PCR product from the first and second rounds, respectively. In the pilot test described in chapter 3 (section 3.2.4), the number of cycles for each round of amplification are detailed in Figure 3.32. In the CRISPRa screen experiment of SAM22 mESCs transduced with the pooled 475 sgRNA lentiviral library, 11 cycles were used in PCR 1, 17 cycles in PCR 2 and 11 cycles in PCR 3. The final amplicon libraries were analysed on a bioanalyzer instrument (Aligent), as shown in Figure 3.33, and multiplexed before sequencing.

Table 2.5– Primer sequences for amplicon sgRNA libraries

| **PCR round** | **Forward primer** | **Reverse primer** |
|---|---|---|
| 1 | TTTCCCATGATTCCTTCATATTTGC | CTACACGACGCTCTTCCGATCT |
| 2 | GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTGTGGAAAGGACGAAACAC | AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC |
| 3 | Nextera XT P7 index in the form of: CAAGCAGAAGACGGCATACGAGAT-[8bp Index]-GTCTCGTGGGCTCGG | AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC |

#### 2.9.1.3 Sequencing

In the MERVL LTR and *Zscan4* CRISPRa experiment (section 3.2.3), each 10X library was sequenced on an Illumina HiSeq4000 lane with 75 cycles for read 1, 75 cycles for read 2 and 8 cycles for i7 sample index. 1,956 E14 mESCs, 2,045 SAM22 mESCs transduced with the non-targeting sgRNA control 1, 2,233 SAM22 mESCs transduced with the MERVL LTR sgRNAs 1 and 3 and 2,362 SAM22 mESCs transduced with the Zscan4 sgRNAs 1 and 3 were captured, based on number of cell barcodes detected, before quality control, with a total of 21,371 genes detected in the dataset.

In the CRISPRa experiment from section 3.2.4, the 10X library was sequenced on an Illumina HiSeq4000 lane with 26 cycles for read 1, 98 cycles for read 2 and 8 cycles for i7 sample index. A total of 3,644 cells were captured based on number of cell barcodes detected, with a total of 19,571 genes detected, before quality control processing. The amplicon sgRNA library was sequenced in a lane of the Illumina Hiseq2500 Rapid Run platform, with 27 cycles for read 1, 267 cycles for read 2 and 8 cycles for i7 sample index.

In the CRISPRa screen experiment of SAM22 mESCs transduced with the pooled 475 sgRNA lentiviral library (chapters 4 and 5), each 10X library was sequenced on an Illumina HiSeq4000 lane with 26 cycles for read 1, 98 cycles for read 2 and 8 cycles for i7 sample index. A total of 114,866 cells were captured for replicate 1, 118,646 cells for replicate 2 and 107,591cells for replicate 3, based on number of cell barcodes detected, which, after merging transduction replicates, resulted in a dataset of 341,103 cells with a total of 23,760 genes detected before quality control processing. The amplicon sgRNA libraries were sequenced across two lanes of the Illumina Hiseq2500 Rapid Run, with 27 cycles for read 1, 267 cycles for read 2 and 8 cycles for i7 sample index.

### 2.9.2 Data processing and analysis

#### 2.9.2.1 Pilot test in section 3.2.3

10X Genomics scRNA-seq data was processed with the default CellRanger v2.1 pipeline (G. X. Y. Zheng et al. 2017) for mapping to the mm10 mouse genome assembly. Gene counts were

further analysed with scanpy (Wolf et al. 2018). To keep only high-quality data and discard potential cell doublets, the following criteria was used to filter cells out: cells with less than 15,000 and more than 40,000 UMI counts; cells with less than 4,000 and more than 6,500 detected genes; and cells with more than 5% of UMI reads coming from mitochondrial genes (Figure 3.20). After this quality control, 1,138 E14 mESCs, 687 SAM22 mESCs transduced with the non-targeting sgRNA control 1, 1,227 SAM22 mESCs transduced with the MERVL LTR sgRNAs 1 and 3, and 899 SAM22 mESCs transduced with the Zscan4 sgRNAs 1 and 3 were retained for analysis.

A gene was considered for downstream analysis if it was detected (UMI count >0) in at least 10 cells that passed the quality control filter across the full dataset. The final dataset consisted of 16,498 genes across 3,951 cells. The number of UMIs for each cell and gene were adjusted by the library size in each cell, dividing by the total number of UMIs per cell. Gene expression levels were obtained as $\log_e$-transformed adjusted UMI counts, scaled by a factor of 10,000. Quantification of MERVL repeat elements was done as described in section 2.9.2.3.5. Differentially expressed genes were estimated with EdgeR, false discovery rate (FDR) <5% (Figure 3.26).

### 2.9.2.2 Pilot test in section 3.2.4

In this test, the whole-transcriptome standard 10X V2 libraries were constructed and sequenced as described in section 2.9.1, but they were not formally processed or analysed. This experiment was primarily used to test and optimise the protocol to perform amplicon sgRNA libraries. In these libraries, the potential sgRNA sequence (nucleotides 24-43 of the read) was compared to the collection of 11 possible sgRNAs and the number of barcodes recovered associated with a sgRNA was calculated (Figure 3.34).

### 2.9.2.3 Pooled screen with 475 sgRNAs

#### 2.9.2.3.1 Data processing, quality control and gene expression quantification

10X Genomics scRNA-seq data was processed with the default CellRanger v2.1 pipeline (G. X. Y. Zheng et al. 2017) for mapping to the mm10 mouse genome assembly. Gene counts were further analysed with scanpy (Wolf et al. 2018). To keep only high-quality data and discard potential cell doublets, the following criteria was used to filter cells out: cells with less than 4,000 and more than 20,000 UMI counts; cells with less than 1,600 and more than 5,000 detected genes; and cells with more than 5% of UMI reads coming from mitochondrial genes (Figure 4.12). After filtering, 109,061 cells were retained from replicate 1, 118,646 cells from replicate 2 and 107,591 cells from replicate 3. Next, cells were assigned a sgRNA using the amplicon sgRNA libraries (see section 2.9.2.3.2) and cells that were not uniquely assigned to one sgRNA were discarded for downstream analysis. This resulted in 71,047 cells in replicate 1, 71,188 cells in replicate 2 and 61,729 cells in replicate 3, which corresponds to 203,894 cells in total across all replicate sets (Figure 4.13).

A gene was considered for downstream analysis if it was detected (UMI count >0) in at least 10 out of the 203,894 cells that passed quality control filtering. The final dataset after quality control consisted of 20,690 genes. The number of UMIs for each cell and gene were adjusted by the library size in each cell, dividing by the total number of UMIs per cell. Gene expression levels were obtained as $\log_e$-transformed adjusted UMI counts, scaled by a factor of 10,000.

For principal component analysis (PCA) (Figures 4.17, 5.2 and 5.4), 965 highly variable genes were selected (Appendices E and F), as implemented in scanpy (Wolf et al. 2018) with minimum mean of 0.01, maximum mean of 5 and minimum dispersion of 0.5. Uniform Manifold Approximation and Projection (UMAP) (Figure 4.17) was also performed on the 965 highly variable genes, as implemented in scanpy (Wolf et al. 2018), with a spread of 3.0.

For mapping dCas9-VP64 and MS2-p65-HSF1 transcripts, reads discarded by CellRanger pipeline for not aligning to the mm10 mouse genome assembly were mapped to an artificial genome assembly containing the integration sequences between the 3’ and 5’ LTR of the

corresponding lentiviral vectors (Figure 3.2), using BWA (version 0.7.17-r1188, default parameters) (Heng Li & Durbin 2009). Reads sharing a UMI and a cell barcode were then collapsed to estimate the number of transcript reads per cell (Figure 4.21).

#### 2.9.2.3.2 Assignment of sgRNAs to cells

Using the amplicon sgRNA libraries, the potential sgRNA sequence (nucleotides 24-43 of the read) was compared to the collection of 475 designed sgRNAs. By taking only exact matches to the white list of sgRNAs, most of the 475 sgRNA sequences were recovered (470-474, variable from one library to another) and 16% of reads on average (15.3%-16.8% for different libraries) were left unassigned to a sgRNA. To correct for PCR and sequencing errors, a minimum edit distance (Levenshtein distance – 4 edits) was allowed between any two sequences of the designed sgRNAs as well as the CROP-sgRNA-MS2 vector sequence surrounding the potential sgRNA in the read. For the reads left unassigned to a sgRNA at this stage, if there was a sgRNA sequence within Levenshtein distance of 1 or 2 and if the 23 nucleotides upstream and 23 nucleotides downstream matched the CROP-sgRNA-MS2 vector sequence with up to 4 edits each, the respective sgRNA was assigned to the read. After this correction procedure, approximately 2% of amplicon reads were left unassigned. Cell barcodes detected in the amplicon libraries were then matched with barcodes detected in the regular 10X Genomics scRNA-seq libraries. Out of the 317,847 cells that passed quality control across the three transduction replicates in the regular 10X libraries, 249,767 cell barcodes were captured in the amplicon sgRNA libraries (85,993 in replicate 1, 86,671 in replicate 2 and 77,103 in replicate 3). A sgRNA was assigned to a cell if more than 90% of all the amplicon reads containing the sgRNA had the same cell barcode, with a standard error of binomial proportion of less than 10% (e.g. more than 8 reads if all the barcodes are associated with the same sgRNA, 13 reads of the same sgRNA if there were more than one sgRNA for a cell barcode, etc.). Table 2.6 illustrates cell numbers and percentages for each assignment in each transduction replicate (see also Figure 4.13):

Table 2.6– Assignment of sgRNA to cells in CRISPRa scRNAseq screen

| Replicate | Total number of cells | Assignment | Number of cells | Percentage (%) |
|---|---|---|---|---|
| 1 | 85,933 | No sgRNA | 397 | 0.46 |
| | 85,933 | Unique sgRNA | 71,047 | 82.62 |
| | 85,933 | Two sgRNAs | 3,028 | 3.52 |
| | 85,933 | Multiple sgRNAs | 11,521 | 13.40 |
| 2 | 86,671 | No sgRNA | 400 | 0.46 |
| | 86,671 | Unique sgRNA | 71,118 | 82.14 |
| | 86,671 | Two sgRNAs | 3,210 | 3.70 |
| | 86,671 | Multiple sgRNAs | 11,873 | 13.70 |
| 3 | 77,103 | No sgRNA | 381 | 0.49 |
| | 77,103 | Unique sgRNA | 61,729 | 80.06 |
| | 77,103 | Two sgRNAs | 3,084 | 4.00 |
| | 77,103 | Multiple sgRNAs | 11,909 | 15.45 |

#### 2.9.2.3.3 Target gene activation

In chapter 4, section 4.2.4, to call activated (higher target gene expression than the basal level in non-targeting sgRNA controls) and not-activated (equal or lower target gene expression than the basal level in non-targeting sgRNA controls) cells for each targeting sgRNA, an empirical p-value was obtained for the expression the corresponding target gene in every cell transduced with such sgRNA after calculating the percentage of cells with a non-targeting sgRNA control that had the same or higher target gene expression. These empirical p-values were then adjusted using Benjamini-Hochberg multiple testing correction. Cells with an FDR of >10% were called

not-activated for the target gene and cells with an FDR of <10% were called activated for the target gene (Figure 4.18 and Appendix C).

#### 2.9.2.3.4 sgRNA off-target effects estimation

To estimate off-target effects for each targeting sgRNA, the sgRNA sequence followed by an NRG PAM (where N can be any nucleotide and R can be A or G), was mapped to the mm10 mouse genome using BWA (version 0.7.17-r1188) (Heng Li & Durbin 2009) with the following parameters: bwa aln -n 5 -l 10 -k 5, which means a maximum of 5 mismatches, a seed length of 10 and a maximum difference of 5 nucleotides in the seed; and bwa samse -n 100, to obtain and report up to 100 alignments per sgRNA. After interpreting the results, sgRNAs were classified in Appendix C as not having or potentially having off-target effects (multi-mapping). Only 9 out of 458 targeting sgRNAs (excluding MERVL LTR sgRNAs) mapped to multiple regions of the genome with up to 5 mismatches.

#### 2.9.2.3.5 Quantification of repeat elements

All occurrences in the genome of repeat sequences from 12 repeat families (LINE-1, LINE-2, ERV1, ERVK, MERVL, Major satellites, Minor satellites, Ribosomal RNA, SINE Alu B1, SINE B2, SINE B4, Telomeric repeats), with each respective genomic locations were downloaded from the UCSC table browser (RepeatMasker, mm10, Nov 2018), concatenated and treated as a reference genome to map the reads discarded by CellRanger pipeline, due to mapping to multiple regions, using SAMtools (Heng Li et al. 2009) and BWA (version 0.7.17-r1188, default parameters) (Heng Li & Durbin 2009). The following number of reads were discarded by CellRanger pipeline in each transduction replicate: 253,330,874 reads in replicate 1, 276,401,843 reads in replicate 2 and 242,863,617 reads in replicate 3, out of which 38,792,331 (15.31%) in replicate 1, 37,285,351 (13.49%) in replicate 2 and 25,837,444 (10.64%) mapped to repeat elements. LINE-2 elements and minor satellite repeats were discarded for downstream analysis due to inefficient mapping (Table 2.7). Reads sharing a UMI and a cell barcode were then collapsed in order to get an estimate of the number of molecules for each repeat family in every cell (Figure 5.6).

Table 2.7– Number of mapped scRNA-seq reads to each repeat family

| Replicate | Repeat family | Number of mapped reads |
|---|---|---|
| 1 | LINE-1 | 15,930,866 |
| | LINE-2 | 56 |
| | ERV1 | 1,049,021 |
| | ERVK | 14,783,904 |
| | MERVL | 1,275,515 |
| | Major Satellites | 305,524 |
| | Minor Satellites | 3 |
| | Ribosomal RNA (rRNA) | 773 |
| | SINE Alu B1 | 398,331 |
| | SINE B2 | 4,583,807 |
| | SINE B4 | 464,024 |
| | Telomeric repeats | 447 |
| 2 | LINE-1 | 14,790,530 |
| | LINE-2 | 86 |
| | ERV1 | 1,159,969 |
| | ERVK | 13,299,950 |
| | MERVL | 1,382,105 |
| | Major Satellites | 318,816 |
| | Minor Satellites | 2 |
| | Ribosomal RNA (rRNA) | 846 |
| | SINE Alu B1 | 457,910 |
| | SINE B2 | 5,341,151 |
| | SINE B4 | 533,479 |

| | | |
|---|---|---|
| | Telomeric repeats | 507 |
| 3 | LINE-1 | 9,769,600 |
| | LINE-2 | 43 |
| | ERV1 | 952,484 |
| | ERVK | 8,865,749 |
| | MERVL | 1,049,799 |
| | Major Satellites | 173,969 |
| | Minor Satellites | 4 |
| | Ribosomal RNA (rRNA) | 551 |
| | SINE Alu B1 | 370,782 |
| | SINE B2 | 4,226,418 |
| | SINE B4 | 427,651 |
| | Telomeric repeats | 394 |

#### 2.9.2.3.6 Multi-omics factor analysis (MOFA) model

A hierarchical Bayesian model, as implemented in an extension of multi-omics factor analysis (MOFA) (Argelaguet et al. unpublished), was trained on two views: first, the set of 965 highly variable protein-coding genes used for PCA and UMAP (see section 2.9.2.3.1), and second, the expression levels of eight repeat families (LINE-1, ERV1, ERVK, MERVL, Major satellites, SINE Alu B1, SINE B2 and SINE B4). Ribosomal RNA (rRNA) and telomeric repeats were excluded from the model due to low detection rate (Figure 5.6). The sgRNA-cell assignment was provided to the model in order to take advantage of group-wise sparsity of the model (Figure 5.5). Upon interpreting the first five factors using their top loadings of variance explained (Appendices G and H, Figures 5.7-5.9), factor 3 was interpreted as a ZGA-like factor due to its top loadings (both protein-coding genes and MERVL repeat) being associated to ZGA.

#### 2.9.2.3.7 Identification of screen hits

MOFA factor 3 was used to reconstruct the expression dataset for each group of cells (grouped by sgRNA expression) to calculate the variance explained by factor 3 for each group. Average fraction of expression variance for protein-coding genes and repeat elements was then used to rank the 460 targeting sgRNAs (Appendix C), with higher percent of variance explained being attributed to sgRNAs with more potency to induce a ZGA-like transcriptional signature. For hit calling, the average and one standard deviation of the fraction of expression variance explained by MOFA factor 3 for cells expressing the fifteen different non-targeting sgRNA controls was estimated as a background rate, and sgRNAs ranking higher than the standard deviation above the mean were considered hits (Figure 5.10, Appendix C). Higher fraction of expression variance explained by MOFA factor 3 correlated with higher factor 3 values and hence higher expression of ZGA-like transcripts (Figures 5.11 and 5.12).

#### 2.9.2.3.8 Differential gene expression analysis

First, for every targeting sgRNA, differential gene expression was assessed between targeted and non-targeted cells transcriptome-wide using a generalised linear model (glm), as implemented in EdgeR, fitted for every gene and a likelihood ratio test to estimate the effect of the targeting sgRNA on the gene's level of expression (Robinson et al. 2010). Using an FDR <10%, few genes were found to be differentially expressed (Appendix C). Consequently, an FDR <100% was applied to rank the top 400 upregulated genes by their FDR in cells expressing each targeting sgRNA compared to non-targeting sgRNA controls. Then, the rank was intersected with a list of known ZGA genes (as described in Appendix I) to assess the enrichment of known ZGA transcripts amongst the differentially expressed genes by each sgRNA (Figure 5.15, Appendix C).

## 2.10 Single-cell NOMe, methylation and transcription-sequencing (scNMT-seq)

### 2.10.1 Preparation of libraries and sequencing

scNMT-seq libraries for *Dppa3* CRISPRa were generated as described in Clark et al. 2018. Briefly, 44 live SAM22 mESCs stably transduced with the Dppa3 sgRNA 379 (Appendix C) and 44 live SAM22 mESCs stably transduced with the non-targeting sgRNA 462 (Appendix C) were single-cell FACS sorted for $DAPI^-$ (see section 2.7) into a GpC methyltransferase-containing mix (M.CviPI, NEB), which was used to label open chromatin by incubation at 37°C for 15 minutes. Subsequently, the RNA and DNA were physically separated using oligo(d)T-conjugated beads. Then, the RNA was processed on the one hand to obtain the transcriptome information, and the DNA was bisulfite-treated and libraries processed on the other hand to obtain chromatin accessibility and DNA methylation information, as described in Clark et al. 2018. Additionally, eight negative controls were included for quality checks: four containing the GpC methylase mix without a cell and four containing a cell and reaction mix but without the GpC enzyme. The multiplexed scRNA-seq libraries were sequenced at 75 bp paired-end on the Illumina NextSeq500 MidOutput and the multiplexed single-cell bisulfite and NOMe libraries were sequenced at 100 bp paired-end on the Illumina HiSeq2500 Rapid Run platform.

### 2.10.2 Data processing, quality control and quantifications

Raw FastQ files of scRNA-seq data were trimmed with Trim Galore (v0.4.1-v0.6.1, default parameters) and mapped to the mouse mm10 genome assembly using Hisat2 (v2.0.5) (D. Kim et al. 2015) with the options –dta –sp1000,1000 –no-mixed –no-discordant. Single-cell bisulfite and NOMe libraries were trimmed with Trim Galore (v0.6.1, default parameters) and mapped to the mouse mm10 genome assembly using Bismark (v0.20.1) (F. Krueger & Andrews 2011) with the –NOMe option.

For scRNA-seq quality control, cells with less than 3,500 genes were discarded, resulting in a dataset of 59 single-cell transcriptomes, 38 expressing the Dppa3 sgRNA 379 and 21

expressing the non-targeting sgRNA control 462 (Figure 4.37). Gene expression was quantified at mRNA level based on Ensembl v96 using the RNA-sequencing quantification pipeline in SeqMonk software (www.bioinformatics.babraham.ac.uk/projects/seqmonk/).

For chromatin accessibility analysis, cells with less than 5,000,000 GpC sites covered were discarded, resulting in 58 cells that passed this quality control in the dataset, 27 expressing the Dppa3 sgRNA 379 and 31 expressing the non-targeting sgRNA control 462. Percentage global accessibility for each cell was calculated as the percentage of methylated GpCs over total GpCs in 10 kilobases windows that had a minimum of 10 GpC sites (Figure 4.38).

For DNA methylation analysis, cells with less than 500,000 CpG sites covered were discarded, resulting in 59 cells that passed this quality control in the dataset, 27 expressing the Dppa3 sgRNA 379 and 32 expressing the non-targeting sgRNA control 462. Percentage methylation for each cell was calculated as the percentage of methylated CpGs over total CpGs in 10 kilobases windows that had a minimum of 10 CpG sites (Figure 4.40).

In the correlation analysis between *Dppa3* promoter accessibility and *Dppa3* expression, only cells that passed quality control criteria for both omics and had at least 10 GpC sites covered in the *Dppa3* promoter region (1.5 kilobases upstream and 0.5 kilobases downstream of the *Dppa3* TSS) were analysed, resulting in a total of 21 cells, 11 expressing the Dppa3 sgRNA 379 and 10 expressing the non-targeting sgRNA control 462 (Figure 4.39). Percentage chromatin accessibility in the *Dppa3* promoter was calculated as the percentage of methylated GpCs over total GpCs. In the correlation analysis between *Dppa3* promoter DNA methylation and *Dppa3* expression, only cells that passed quality control criteria for both omics were analysed, resulting in a total of 30 cells, 17 expressing the Dppa3 sgRNA 379 and 13 expressing the non-targeting sgRNA control 462 (Figure 4.41). Percentage DNA methylation in the *Dppa3* promoter was calculated as the percent of methylated CpG over total CpG, without a thereshold for a minimum number of observations.

## 2.11 Data processing and analysis of published sequencing datasets

The RNA-sequencing datasets from Xue et al. 2013, Eckersley-Maslin et al. 2019, Rulands et al. 2018, Deng et al. 2014 and Barisic et al. 2019 were processed and quantified as described in section 2.8.2. Transcriptional variability between single cells from Rulands et al. 2018 was quantified as follows: highly variable genes were calculated considering the over-dispersion observed in scRNA-seq data (Brennecke et al. 2013); a score of variability per gene was calculated by fitting the squared coefficient of variation as a function of the mean normalised counts and then calculating the distance to a rolling average (window size=100); to minimize the bias due to lowly expressed genes, only genes with a mean greater than 10 were used.

For processing of ATAC-sequencing (King et al. 2018; Dongwei Li et al. 2017), DNase-sequencing (ENCODE) and ChIP-sequencing (Hernandez et al. 2018; Bilodeau et al. 2009; Neri et al. 2017) data, raw FastQ files were trimmed with Trim Galore (v0.6.1, default parameters) and mapped to the mouse mm10 genome assembly using Bowtie2 (v2.3.3, default parameters) (Langmead & Salzberg 2012). Hits were filtered to remove mappings with MAPQ scores <20. $Log_2$ normalised read counts were quantified over promoters defined as the region 1 kilobase upstream of the gene TSS.

For processing of whole-genome bisulfite sequencing libraries from Ficz et al. 2013, Habibi et al. 2013, Milagre et al. 2017 and Berrens et al. 2017, raw FastQ files were trimmed with Trim Galore (v0.6.1, default parameters) and mapped to the mouse mm10 genome assembly using Bismark (v0.20.1, default parameters) (F. Krueger & Andrews 2011). CpG methylation calls were extracted and analysed using SeqMonk software (www.bioinformatics.babraham.ac.uk/projects/seqmonk/). Methylation over promoters, defined as the region 1 kilobase upstream of the gene TSS, was calculated as the percentage of methylated CpGs over total CpGs considering only CpG sites covered by at least 10 reads and only promoters with at least 10 CpGs covered.

## 2.12 Other software and data representation

For screen candidate selection in section 4.2.1, mouse proteins associated with nucleic acid binding and transcription factor activities were extracted from the PANTHER (http://pantherdb.org) (Mi et al. 2019) protein classes PC00171 and PC00218. Motif enrichment and discovery analysis on the promoters of activated and not-activated genes, as described in section 4.2.4, was performed using AME and DREME tools from the MEME suite (http://meme-suite.org/index.html) (Bailey et al. 2009). Gene ontology analysis for gene loadings of PCA components (Appendix F) and MOFA factors (Appendix H), as described in section 5.2.1, was performed with Gorilla (http://cbl-gorilla.cs.technion.ac.il) (Eden et al. 2009). Graphs and illustrations were generated with RStudio, SeqMonk, GraphPad Prism, Microsoft Excel and Illustrator software.

## 2.13 Sequencing data availability

The following sequencing data has been deposited in GEO (GSE135622): bulk RNA-sequencing of SAM22 and E14 mESCs; scRNA-seq of E14 and CRISPRa of MERVL LTRs and *Zscan4*; scRNA-seq of CRISPRa screen data; and bulk RNA-sequencing of CRISPRa and cDNA overexpression of *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1*.

# Chapter 3 Development of CRISPR activation at single-cell RNA-sequencing resolution

## 3.1 Background and summary

The aim of my thesis is to comprehensively understand the regulation of ZGA, a crucial transcriptional event that occurs early after fertilisation (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018; Svoboda 2017; Jukam et al. 2017; Yartseva & Giraldez 2015). However, screening in early mouse embryos is not feasible at a high-throughput level due to the scarcity of material, maternal stores of proteins and complex manipulation techniques required. Recent studies have shown that a major ZGA-like state can be mimicked in mouse embryonic stem cells (mESCs) (Macfarlan et al. 2012; Zalzman et al. 2010; Bošković et al. 2014; Ishiuchi et al. 2015; Akiyama et al. 2015; Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018). These cells are termed 2C-like cells as they partially resemble the two-cell embryo. Consequently, they represent an ideal system for *in vitro* screening and have been previously used to identify regulators of ZGA (Rodriguez-Terrones et al. 2018; Eckersley-Maslin et al. 2019; X. Fu et al. 2019; Y.-L. Yan et al. 2019).

Most of the regulators of 2C-like cells identified to date are repressors of this state (Macfarlan et al. 2011; Rowe et al. 2010; Macfarlan et al. 2012; Ishiuchi et al. 2015; Percharde et al. 2018; Rodriguez-Terrones et al. 2018; Choi et al. 2017; Maksakova et al. 2013; Storm et al. 2014; Schoorlemmer et al. 2014; Hisada et al. 2012; X. Fu et al. 2019; Y.-L. Yan et al. 2019) (see section 1.2.3) and, therefore, cannot explain how ZGA is triggered in the embryo. The identification of positive inducers of ZGA is more relevant given the transcriptionally inactive state prior to ZGA. During the course of this dissertation, DUX was reported as positive crucial regulator of ZGA in mouse and human ESCs (Hendrickson et al. 2017; De Iaco et al. 2017; Whiddon et al. 2017). However, *Dux* is only expressed at the onset of ZGA (Hendrickson et

al. 2017; De Iaco et al. 2017), which raises the question of what are the maternally-deposited proteins that activate *Dux*. Furthermore, recent studies have shown that mouse and human embryos that lack maternal and/or zygotic *Dux* are not heavily compromised (Iaco et al. 2019; Z. Chen & Yi Zhang 2019; Vuoristo et al. 2019), indicating that DUX is a non-essential synchronizer of ZGA and suggesting redundancy in the regulation of ZGA.

All of these positive and negative regulators were identified using reporter constructs of two well-characterised ZGA transcripts: MERVL (Kigami 2002; Peaston et al. 2004; Svoboda et al. 2004; Macfarlan et al. 2012) and *Zscan4* gene cluster (Falco et al. 2007; Ishiguro, Nakatake, et al. 2016; Zalzman et al. 2010). Even though the single and double positive populations of mESCs for these reporters recapitulate well the two-cell embryo at the time of major ZGA, their transcriptome and epigenome are not identical (reviewed in Ishiuchi & Torres-Padilla 2013; Eckersley-Maslin, Alda-Catalinas & Reik 2018). Consequently, the use of whole transcriptome read-out in a mESC screen could benefit from potentially identifying regulators that trigger a wider ZGA response than the one observed in 2C-like cells. Moreover, such a screen design requires a method to overexpress those potential ZGA regulators in a high-throughput manner, reasoning that their upregulation could induce a ZGA signature. On the contrary, screening by knocking-out or downregulating the expression of candidates would only identify repressors of the 2C-like state, which might act *in vivo* by repressing ZGA transcripts after the two-cell stage, but could not identify regulators that trigger ZGA.

CRISPRa is a potent tool for selective transcriptional activation of endogenous genes. Using sgRNAs targeted upstream of TSSs together with a catalytically-inactive Cas9 (or dead Cas9, dCas9) and transcriptional co-activators, RNA polymerases can be recruited to gene promoters and activate or increase transcription of the endogenous locus (Mali, Aach, et al. 2013; Gilbert et al. 2013; A. W. Cheng et al. 2013; Perez-Pinera et al. 2013; Maeder et al. 2013; Gilbert et al. 2014; Tanenbaum et al. 2014; Konermann et al. 2015; Chakraborty et al. 2014; Chavez et al. 2015; Chavez et al. 2016; Rajagopal et al. 2016, see section 1.4.1). The CRISPRa "synergistic activator mediator" (SAM) uses three trans-activators to induce robust transcription: on the one hand, the tetrameric VP16 transcription activation domain (or VP64) is fused to dCas9; on the other hand, the nuclear factor NF-kappa-β p65 subunit (p65) is fused together with the heat shock factor-1 protein (HSF1) to the RNA binding protein MS2 (Figure

3.1). The MS2-p65-HSF1 fusion protein is recruited to the sgRNA-dCas9-VP64 complex through MS2 loops introduced in the sgRNA scaffold sequence (Konermann et al. 2015) (Figure 3.1).

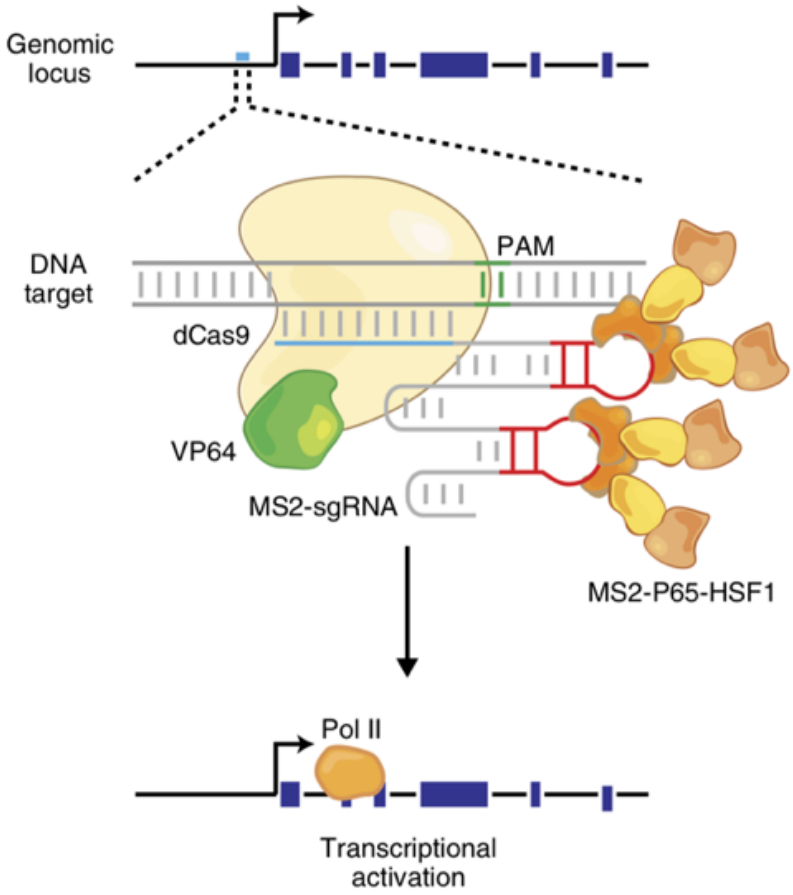



**Figure 3.1– Schematic of CRISPRa SAM method**

Schematic showing the components and function of CRISPRa method SAM. The sgRNA (grey) is targeted upstream of the TSS of the gene of interest and its scaffold contains two MS2 loops (red); the catalytically-inactive or dead Cas9 (dCas9) (yellow) is fused to the trans-activation domain VP64 (green) and the MS2 RNA binding protein (orange) to the trans-activators p65 and HSF1 (yellow and orange). MS2 RNA binding proteins bind to MS2 loops as dimers and therefore, four copies of p65-HSF1 are recruited to the dCas9-VP64/sgRNA complex. Recruitment of these three components to the promoter region of a gene (or upstream of the TSS- depicted by an arrow) induces recruitment of polymerase II (Pol II) and, consequently, transcriptional gene activation. Image from Joung et al. 2017.

Pooling multiple sgRNAs targeting different genes into a single lentiviral library that can be transduced at low MOI into cells allows the introduction of multiple perturbations in a single pooled experiment while having a unique perturbation in every cell (Gilbert et al. 2014;

Konermann et al. 2015; Horlbeck et al. 2016; Sanson et al. 2018, see section 1.4.2). Transcriptomic analysis of single cells can consequently allow the identification of the sgRNA expressed in every cell while having a complex functional interrogation of the targeted gene (see section 1.5). Although CRISPR KO and CRISPRi have been previously combined with single-cell transcriptomic analysis (Jaitin et al. 2016; Dixit et al. 2016; Adamson et al. 2016; Datlinger et al. 2017; S. Xie et al. 2017; Replogle et al. 2018; Genga et al. 2019; Gasperini et al. 2019), CRISPRa screens have been limited to simpler read-outs.

In this chapter, I generated, optimised and characterised the tools necessary to perform a CRISPRa screen at single-cell transcriptomic resolution in mESCs. This characterisation was done with the aim of subsequently applying the method to target multiple maternal epigenetic and transcriptional regulators and uncover new factors that induce a ZGA-like signature in mESCs. First, I generated a clonal mESC line expressing dCas9-VP64 and MS2-p65-HSF1. Next, I tested CRISPRa in mESCs by targeting MERVL and the gene cluster *Zscan4*. I initially tested and characterised the system using qPCR, bulk RNA-sequencing, western-blot and immunofluorescence, and then performed scRNA-seq experiments using the 10X Genomics platform. These tests revealed that CRISPRa works effectively in mESCs and that targeting MERVL and *Zscan4* induces a ZGA-like signature that can be detected both by bulk and single-cell RNA-sequencing measurements. Lastly, I envisaged and validated a strategy to simultaneously capture sgRNA sequence information together with the whole transcriptome in single-cells by scRNA-seq.

## 3.2 Results

### 3.2.1 Generation of a clonal mouse embryonic stem cell line for CRISPR activation

In order to generate a stable mESC line expressing the CRISPRa SAM components, dCas9-VP64 and MS2-p65-HSF1 (Figure 3.2), I transduced lentiviral constructs expressing these two fusion proteins into E14 mESCs. After two weeks of selection with the antibiotics blasticidin and hygromycin, which were respectively cloned in the integration sites of the corresponding

lentiviral constructs, I manually picked 18 individual colonies, expanded and harvested them for DNA and RNA isolation (see Materials and Methods). I also harvested a non-clonal population by pooling the remnant colonies that were not picked.

Using genomic DNA, I performed a PCR with primers specific for dCas9-VP64 and MS2-p65-HSF1 (see Materials and Methods). After running the resulting PCR products in an agarose gel, I observed that most clones were positive for both integrations, while the non-clonal population was positive for MS2-p65-HSF1 but very weak for dCas9-VP64 integration (Figure 3.3), suggesting that clonal selection improves stable integration of exogenous inserts.

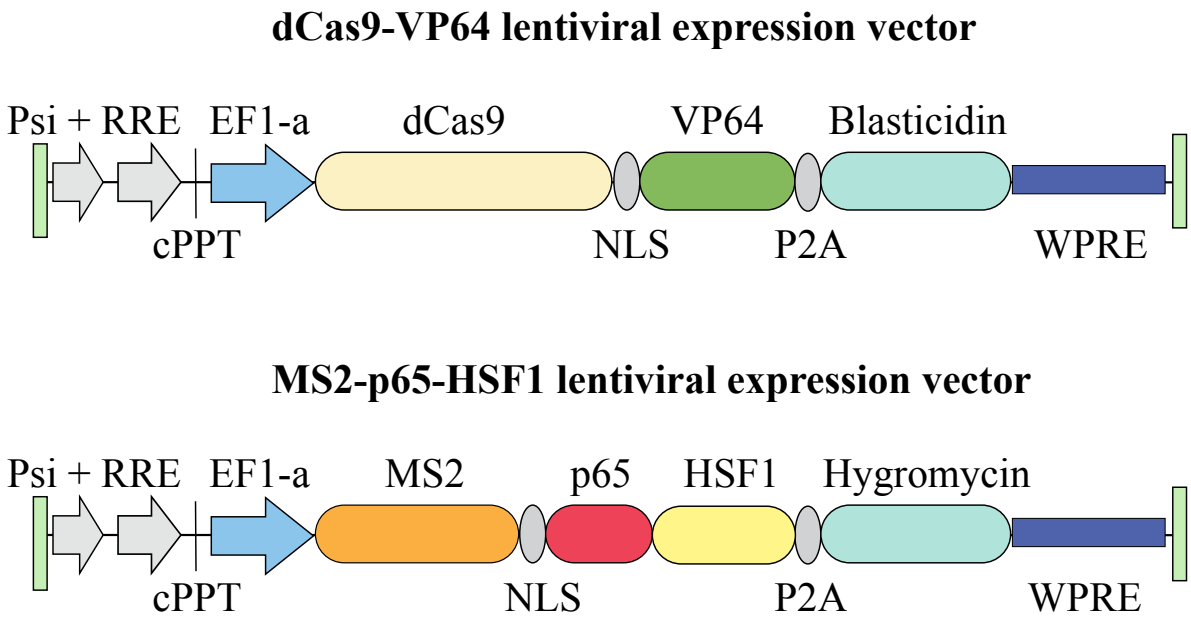



**Figure 3.2– Schematic representation of dCas9-VP64 and MS2-p65-HSF1 lentiviral constructs**

The integration site of both dCas9-VP64 (top panel) and MS2-p65-HSF1 (bottom panel) lentiviral constructs contains a retroviral psi packaging element, a rev response element (RRE), a central polypurine tract (cPPT) and a Pol II elongation factor-1 alpha (EF1α) promoter. In the case of the dCas9-VP64 construct (top panel), EF1α drives the expression of a catalytically inactive or dead Cas9 (dCas9) linked to the trans-activation domain VP64 via a nuclear localisation signal (NLS); a P2A self-cleaving peptide is placed between the dCas9-VP64 fusion and the antibiotic selection marker blasticidin, followed by a Woodchuck hepatitis virus posttranscriptional regulatory element (WPRE). In the MS2-p65-HSF1 construct (bottom panel), EF1α drives the expression of the RNA binding protein MS2 linked to the trans-activation domains p65 (nuclear factor NF-kappa-β p65 subunit) and HSF1 (heat shock factor-1) via an NLS; a P2A self-cleaving peptide is placed between the MS2-p65-HSF1 fusion and the antibiotic selection marker hygromycin, followed by WPRE.

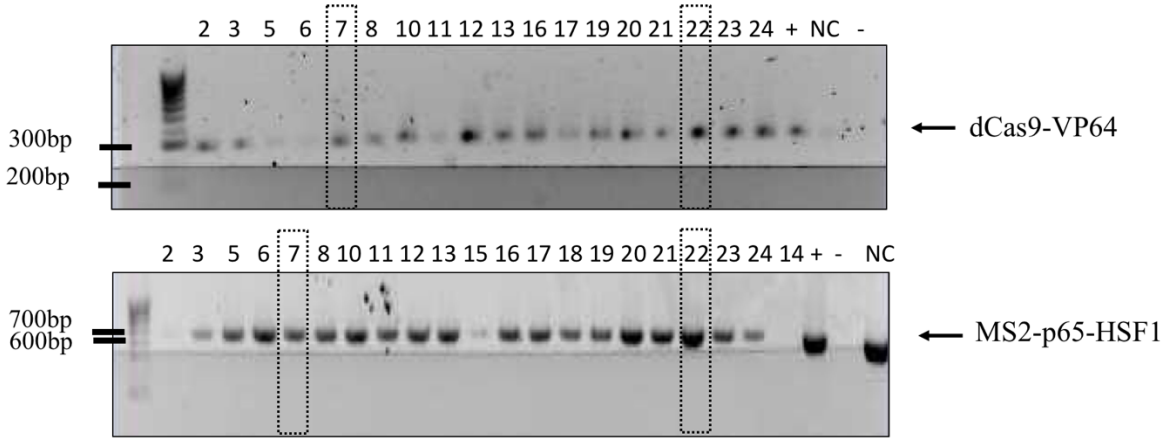



**Figure 3.3– Genomic integrations of SAM embryonic stem cell clones**

Agarose gel electrophoresis showing the PCR products of dCas9-VP64 (top panel) and MS2-p65-HSF1 (bottom panel) after specific amplification from genomic DNA of the SAM mESCs clones. The expected band size for a clone expressing dCas9-VP64 is 275bp and for MS2-p65-HSF1 is 672 bp. The input material for PCR positive controls (+) was plasmid DNA of dCas9-VP64 construct (top panel) or MS2-p65-HSF1 construct (bottom panel). The PCR negative controls (-) included all PCR reagents but DNA. The ladder is HyperLadder IV (Biolane, BIO-33029). Dotted boxes highlight clones 7 and 22 that were picked for further characterisation. NC: non clonal.

To further characterise the transcriptome of these cell lines and confirm expression of dCas9-VP64 and MS2-p65-HSF1, I selected clones SAM7 and SAM22 for RNA-sequencing analysis and compared their transcriptomics to the parental cell line E14. Both clones showed high expression of dCas9-VP64 and MS2-p65-HSF1 and presented a largely unchanged transcriptome compared to E14 mESCs (Figure 3.4). Clone SAM7 had 13 differentially expressed genes (DEGs) and clone SAM22 showed 5 DEGs; these transcriptomic changes are probably not due to the expression of dCas9-VP64 and MS2-p65-HSF1 but rather reflect the expected variation generated during cell culture as well as technical variation of library preparation and sequencing, since two different cultures and RNA-sequencing libraries of E14 mESCs sequenced at the same time showed a similar number of DEGs (Figure 3.4). In conclusion, expression of SAM CRISPRa machinery has minimal effects on gene expression. Due to the lower number of DEGs, I selected clone SAM22 to carry all further experiments presented in this dissertation.

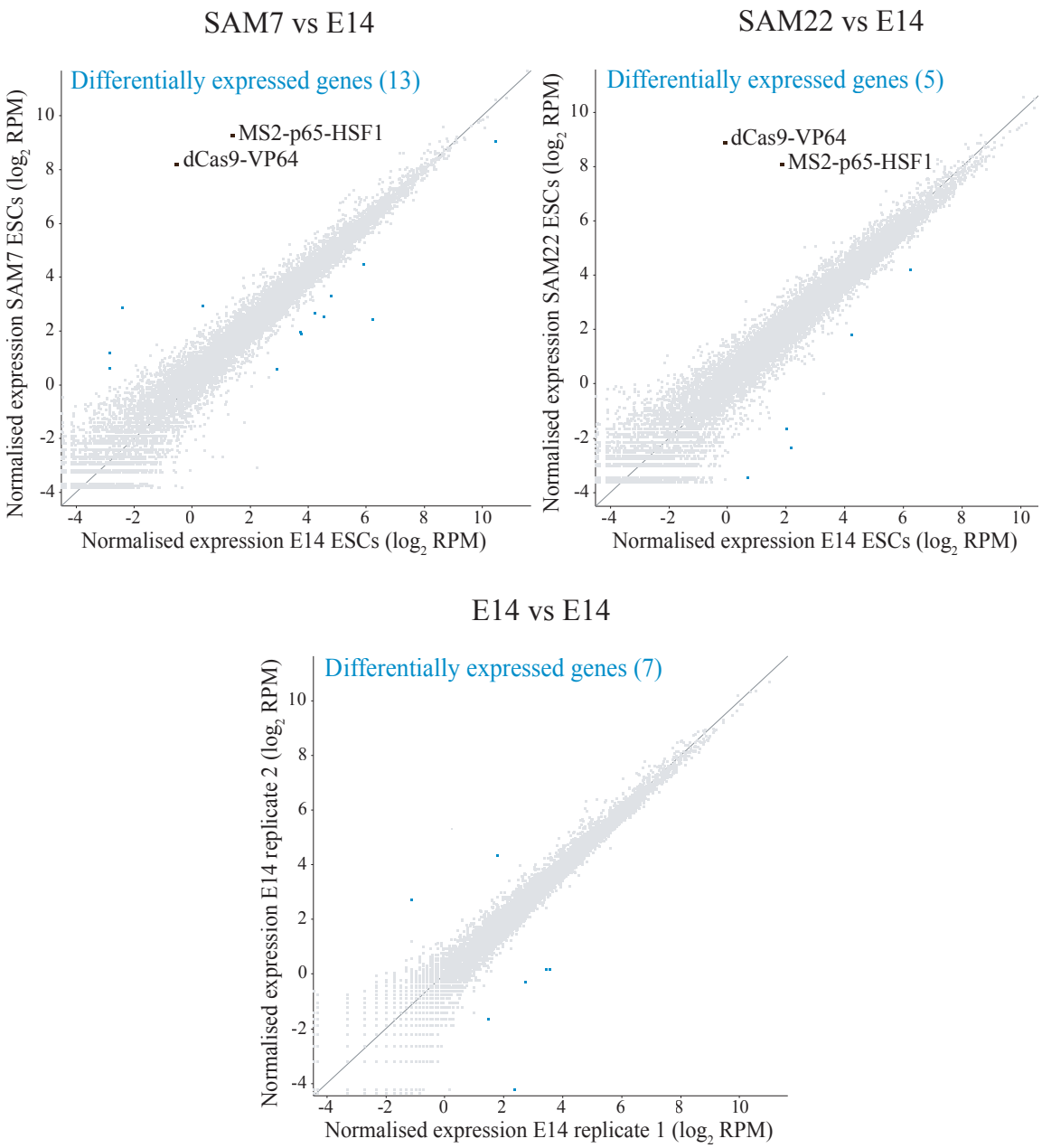



**Figure 3.4– Transcriptome of SAM clones 7 and 22**

Scatterplots showing normalised gene expression in $\log_2$ reads per million (RPM) of SAM7 clone against its parental cell line E14 (top left panel), SAM22 clone against E14 (top right panel) and two replicates of E14s (bottom panel) analysed by RNA-sequencing, highlighting dCas9-VP64 and MS2-p65-HSF1 transcripts in black and differentially expressed genes (DEGs) in blue. The transcriptome analysis was done with two replicates of SAM7, two replicates of SAM22 and three replicates of E14. DEGs were determined using DESeq2 (FDR<5%), EdgeR (FDR<5%) and intensity difference filter (FDR<5%), with the high-confidence DEGs defined as the intersection between the three statistical tests and a $\log_2$ fold change >2.

### 3.2.2 CRISPR activation of MERVL and Zscan4

I next chose to target the endogenous retrovirus MERVL and the gene cluster *Zscan4* to test and optimise my CRISPRa system in mESCs. These transcripts are specifically expressed in mouse two-cell embryos at the time of ZGA and in 2C-like cells (Macfarlan et al. 2012; Zalzman et al. 2010; Eckersley-Maslin et al. 2016). As previously described in Konermann et al. 2015, I designed sgRNAs to target the 200 bp window upstream of the gene TSS for optimal CRISPRa efficiency. I designed and cloned into the sgRNA(MS2)_puro backbone two sgRNAs targeting MERVL long terminal repeats (LTRs) (MERVL sgRNA 1 and MERVL sgRNA 3) and two targeting *Zscan4* cluster promoters (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) as well as a non-targeting sgRNA (see Materials and Methods and Appendix C for sgRNA sequences). Given the high sequence similarity of the genes within the *Zscan4* cluster (*Zscan4b, Zscan4c, Zscan4d, Zscan4e, Zscan4f*), a single sgRNA can target more than one gene within the cluster, specifically, Zscan4 sgRNA 1 targeted the 200 bp upstream window of the TSSs of *Zscan4b, Zscan4c, Zscan4d* and *Zscan4f* whereas Zscan4 sgRNA 3 targeted the promoters of *Zscan4c* and *Zscan4f* (see Appendix C).

After transducing individually each sgRNA into SAM22 mESCs in triplicate and selecting the cells with puromycin, I analysed target gene activation by qPCR six days after transduction. Importantly, the qPCR primers used for *Zscan4* were designed to amplify all transcripts in the cluster. This experiment demonstrates that all four sgRNAs induced at least a 3-fold upregulation of their target loci compared to a non-targeting control (Figure 3.5). Unsurprisingly, sgRNAs targeted to MERVL LTRs also induced *Zscan4* activation and *vice versa*, suggesting that both belong to the same transcriptional network and are inter-regulated.

In order to prove this observation and to check whether RNA upregulation translated into an increase in protein levels, I extracted protein from these samples and analysed ZSCAN4 expression by western-blot. ZSCAN4 protein was highly increased in SAM22 mESCs transduced with *Zscan4* promoter-targeted sgRNAs and also, to a lesser extent, with a MERVL LTR-targeted sgRNA (Figure 3.6). Immunofluorescence analysis of ZSCAN4 protein confirmed these observations, showing that ZSCAN4 was only expressed in a small percentage of cells at basal levels in the non-targeting sgRNA control, in agreement with previous studies

(Zalzman et al. 2010; Eckersley-Maslin et al. 2016), and CRISPRa induced expression in a larger number of cells (Figure 3.7). It is however interesting to note that not all cells transduced with *Zscan4* or MERVL sgRNAs, and selected with puromycin for sgRNA expression, showed ZSCAN4 protein expression, which raises some questions that are discussed later on in this chapter (see section 3.3). Nevertheless, I can conclude that the CRISPRa SAM method in mESCs not only activates transcription of endogenous loci but also leads to corresponding protein overexpression.

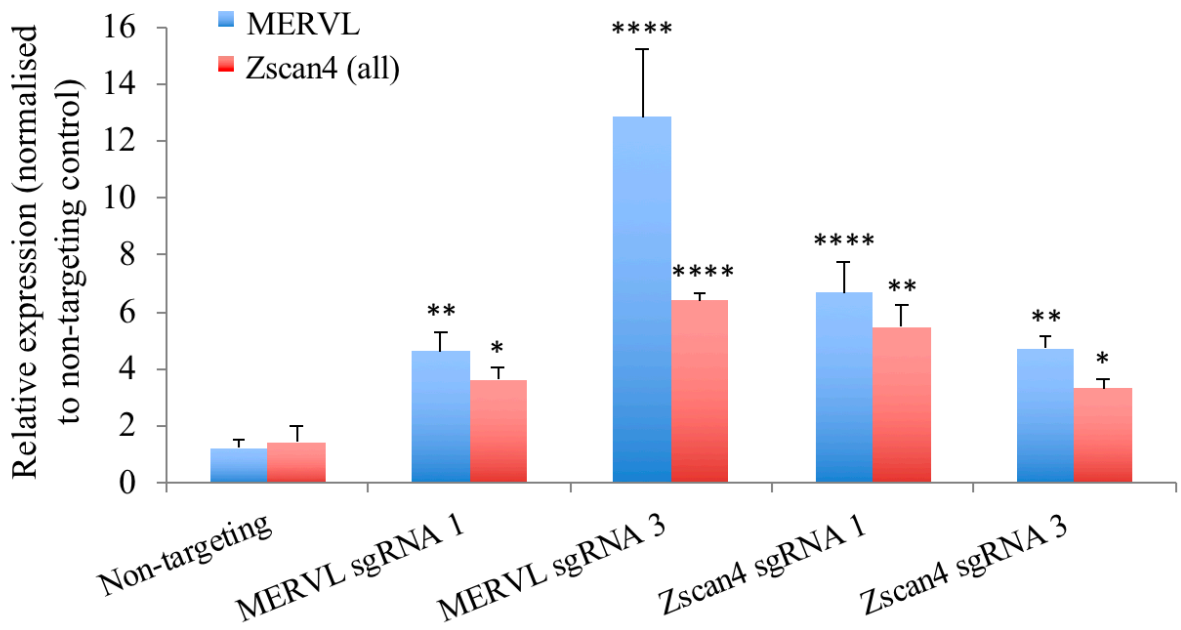



**Figure 3.5– MERVL and *Zscan4* expression induced by CRISPRa, analysed by qPCR**

Analysis of MERVL (blue) and *Zscan4* (red) relative RNA levels by qPCR in SAM22 mESCs transduced with a non-targeting sgRNA control, with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3). RNA was harvested 6 days post-sgRNA transduction and puromycin selection. The experiment was performed in triplicate and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to non-targeting control are reported (homoscedastic two-tailed t-test: ****: p-value < 0.0001, **: p-value <0.01, * p-value <0.5).

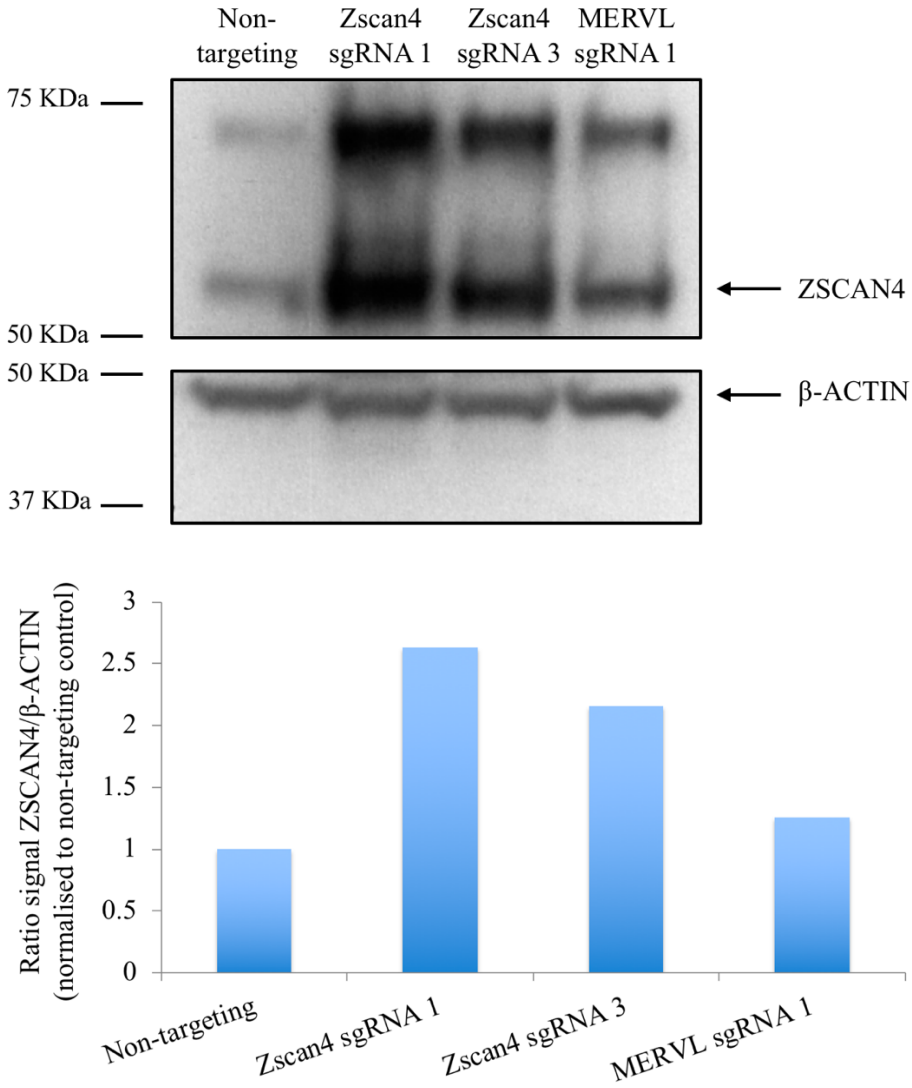



**Figure 3.6– Western-blot of ZSCAN4 after CRISPRa of *Zscan4* and MERVL**

Western-blot showing ZSCAN4 protein (~65KDa) levels in SAM22 mESCs transduced with a non-targeting sgRNA control, Zscan4 sgRNA 1, Zscan4 sgRNA 3 and MERVL sgRNA 1. An uncharacterised ZSCAN4 band was also observed at ~ 72KDa. β-ACTIN antibody was used as a loading control. Protein extracts were harvested 15 days after sgRNA transduction and puromycin selection. The experiment was done once. Quantification of the ~57 KDa Zscan4 bands corrected by the intensity of the corresponding β-actin band and normalised to the non-targeting control is shown in the bottom panel.

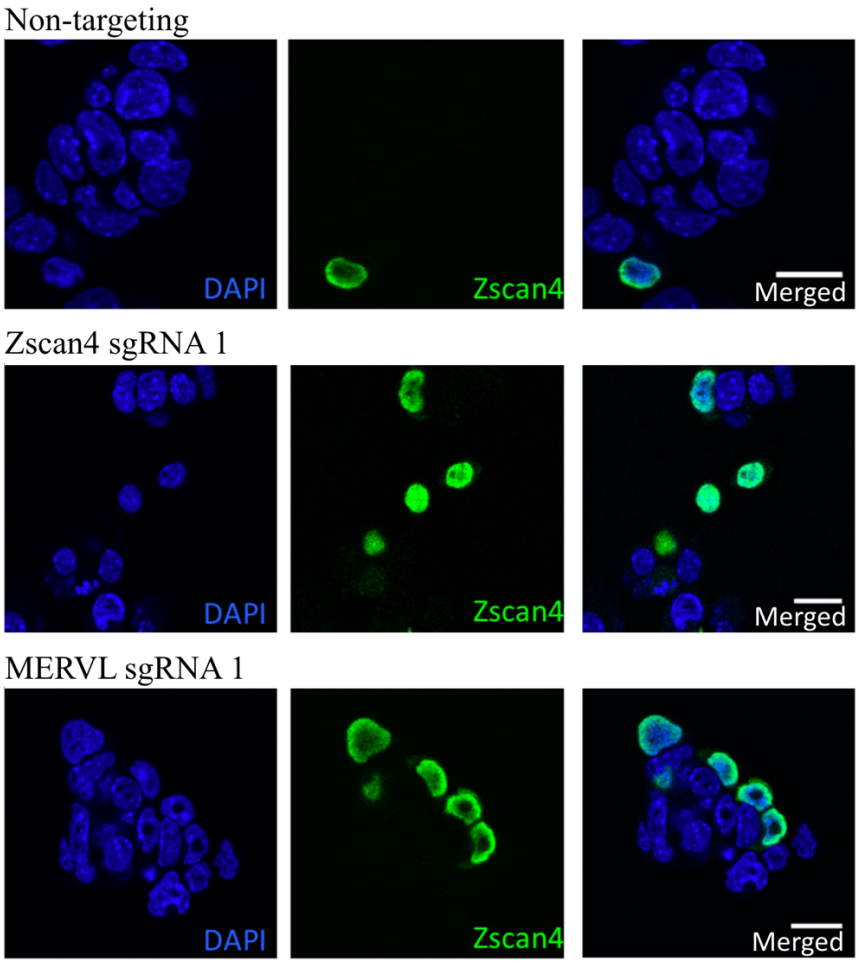



**Figure 3.7– Immunofluorescence of ZSCAN4 after CRISPRa of *Zscan4* and MERVL**

Representative images of SAM22 mESCs transduced with a non-targeting sgRNA control, Zscan4 sgRNA 1 and MERVL sgRNA 1, and stained for ZSCAN4 protein (green) and DNA stained with DAPI (blue). The cells were seeded on coverslips 15 days after sgRNA transduction and puromycin selection and cultured for an extra day before fixation. Images show a single Z-plane. Scale bar = 20 μm.

2C-like mESCs are characterised transcriptionally by the activation of a network of genes driven by MERVL LTR elements, similarly to ZGA in two-cell embryos (Kigami 2002; Peaston et al. 2004; Macfarlan et al. 2012; Y. Huang et al. 2017; Ishiuchi et al. 2015; Eckersley-Maslin et al. 2016). To investigate whether *Zscan4* and MERVL LTR CRISPR-activated cells resembled a 2C-like or ZGA-like state, I initially analysed the expression of a panel of ZGA markers by qPCR, showing that both MERVL-LTR and Zscan4 sgRNAs induced activation of these markers, whereas the pluripotency genes *Pou5f1* (also known as *Oct4*) and *Nanog* did not show any significant change (Figure 3.8).

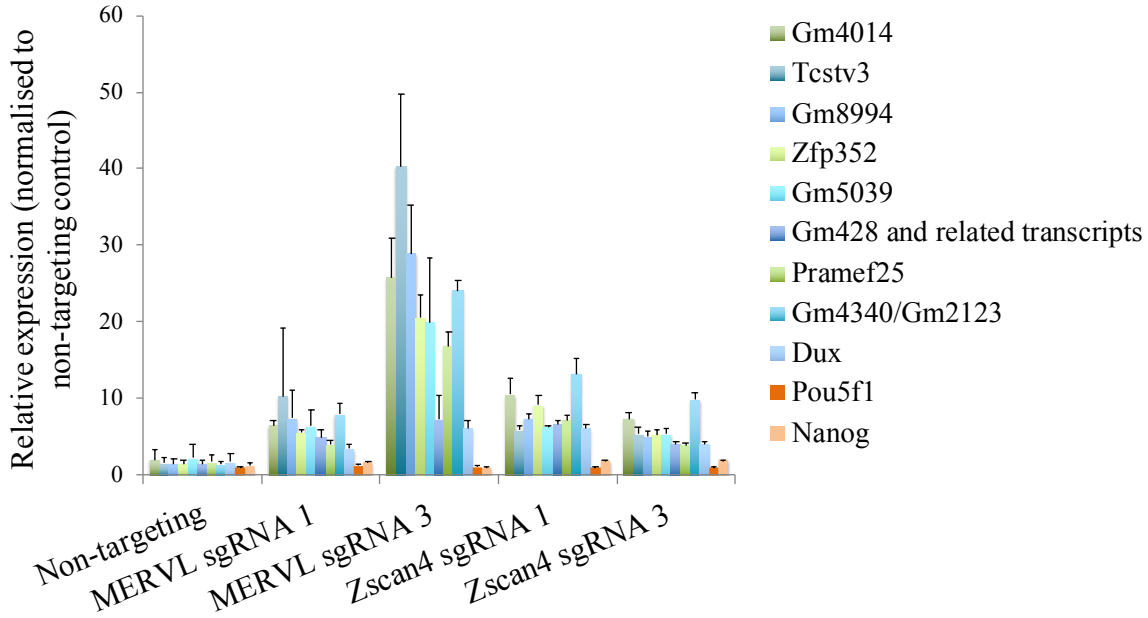



**Figure 3.8– ZGA markers, *Pou5f1* and *Nanog* expression after CRISPRa of MERVL and Zscan4, analysed by qPCR**

Analysis of relative mRNA levels of a panel of ZGA markers (green-blue), *Pou5f1* (*Oct4)* (dark orange) and *Nanog* (light orange) by qPCR in SAM22 mESCs transduced with a non-targeting sgRNA control, with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3). RNA was harvested 6 days post-sgRNA transduction and puromycin selection. The experiment was performed in triplicate and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to non-targeting control were assessed by homoscedastic two-tailed t-test. For figure simplicity, */**/***/**** are not depicted; all transcripts except Pou5f1 and Nanog were significantly upregulated in the four experimental conditions compared to the non-targeting control, with a p-value < 0.05.

To achieve a more comprehensive understanding of the transcriptional changes triggered by MERVL and *Zscan4* CRISPRa, I performed RNA-sequencing analysis on SAM22 mESCs transduced with MERVL LTR or Zscan4 sgRNAs. I first interrogated target gene activation and observed that all genes within the *Zscan4* cluster were activated upon transduction with both MERVL LTR and Zscan4 sgRNAs (Figure 3.9). Interestingly, some sgRNAs were more efficient than others, in particular, Zscan4 sgRNA 1 induced a strong upregulation of *Zscan4b*, *Zscan4c*, *Zscan4d*, *Zscan4f* and, to a lesser extent of *Zscan4e*. This is consistent with the design of the sgRNA (Appendix C) and with the results observed for ZGA marker upregulation by qPCR, where Zscan4 sgRNA 1 was a slightly stronger inductor than Zscan4 sgRNA 3 (Figures 3.8 and 3.9).

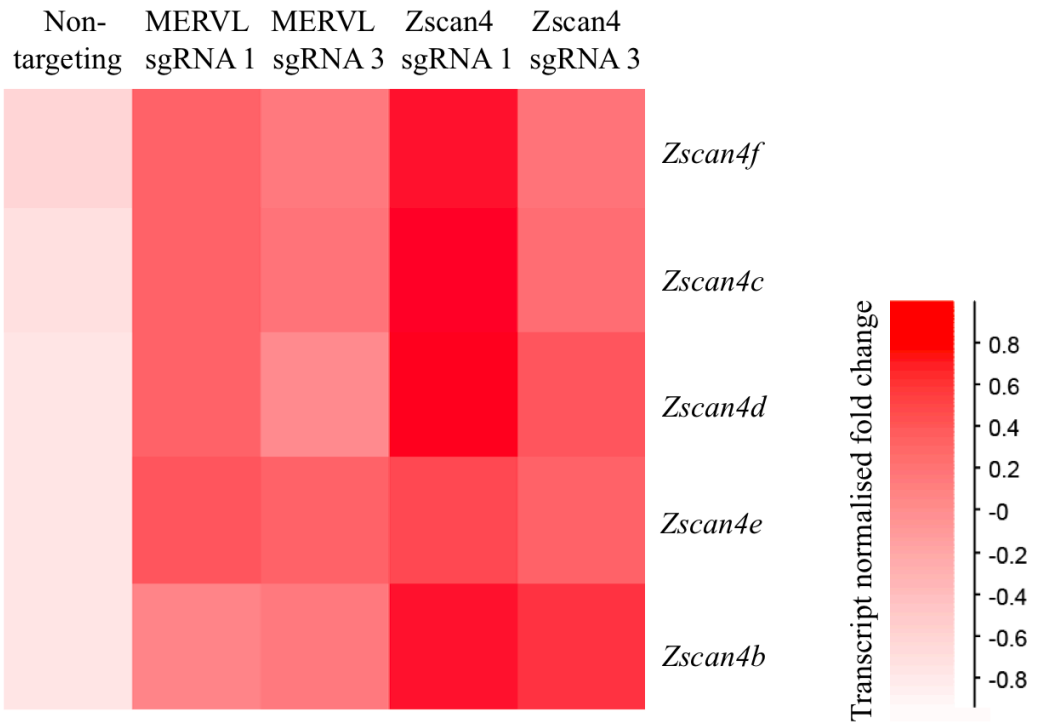



**Figure 3.9– Heatmap of *Zscan4* genes expression after CRISPRa of MERVL and *Zscan4*, analysed by RNA-sequencing**

Heatmap showing transcript normalised expression, scaled per transcript, of *Zscan4f*, *Zscan4c*, *Zscan4d*, *Zscan4e* and *Zscan4b* in SAM22 mESCs transduced with a non-targeting sgRNA control, with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3), analysed by RNA-sequencing. RNA was harvested 6 days post-sgRNA transduction and puromycin selection. The experiment was performed in triplicate.

In order to analyse MERVL target upregulation by RNA-sequencing, the data was re-mapped to a custom-built repeat genome, allowing analysis of different repeat element families. As expected from the qPCR results, the four sgRNAs induced strong upregulation of MERVL elements (Figure 3.10). Interestingly, MERVL sgRNA 3, the strongest sgRNA for MERVL activation as indicated by qPCR analysis (Figure 3.5 and 3.8), as well as the strongest Zscan4 sgRNA (Zscan4 sgRNA 1), also upregulated LINE-1 elements (Figure 3.10). This is consistent with previous studies showing that LINE-1 expression is activated during ZGA (Beraldi et al. 2006; Jachowicz et al. 2017; Percharde et al. 2018; Eckersley-Maslin et al. 2019; De Iaco et al. 2019). Other repeat families, such as ERV1, MaLR or SINE B2, were significantly upregulated by one or more sgRNAs compared to the non-targeting sgRNA control, however, the absolute

change in the percentage of total sequencing reads was very small (Figure 3.10), making it difficult to draw any conclusions on the biological significance.

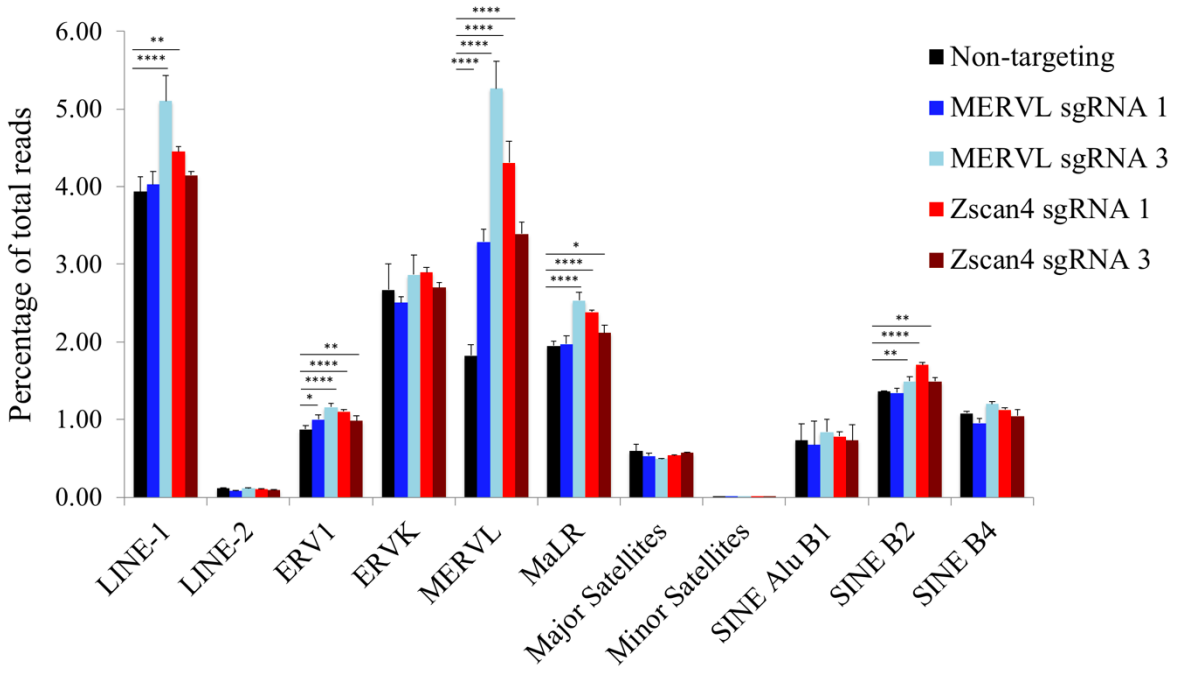



**Figure 3.10– Expression repeat families after CRISPRa of MERVL and *Zscan4*, analysed by RNA-sequencing**

Expression of different repeat families analysed by RNA-sequencing and plotted as percentage of total reads in the library, in SAM22 mESCs transduced with a non-targeting sgRNA control (black), with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1- blue, and MERVL sgRNA 3- light blue) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1- red and Zscan4 sgRNA 3- brown). RNA was harvested 6 days post-sgRNA transduction and puromycin selection. The experiment was performed in triplicate. Data is shown as mean plus standard deviation. Statistically significant differences to non-targeting sgRNA control are reported (homoscedastic two-tailed t-test, ****: p-value < 0.0001, ***: p-value <0.001, **: p-value <0.01, * p-value <0.5, absence of starts: not significant).

Given that MERVL LTR elements can act as functional promoters driving the expression of hundreds of chimeric transcripts (Macfarlan et al. 2012; Y. Huang et al. 2017; Franke et al. 2017), I analysed transcriptional activation of 10 kilobase (kb) probes downstream of MERVL LTR elements. Read quantification of these probes showed that transcription was increased downstream of MERVL LTR elements in cells targeted with MERVL LTR and Zscan4 sgRNAs (Figure 3.11), reinforcing the idea that ZSCAN4- and MERVL-driven transcriptional programs are directly inter-regulated.

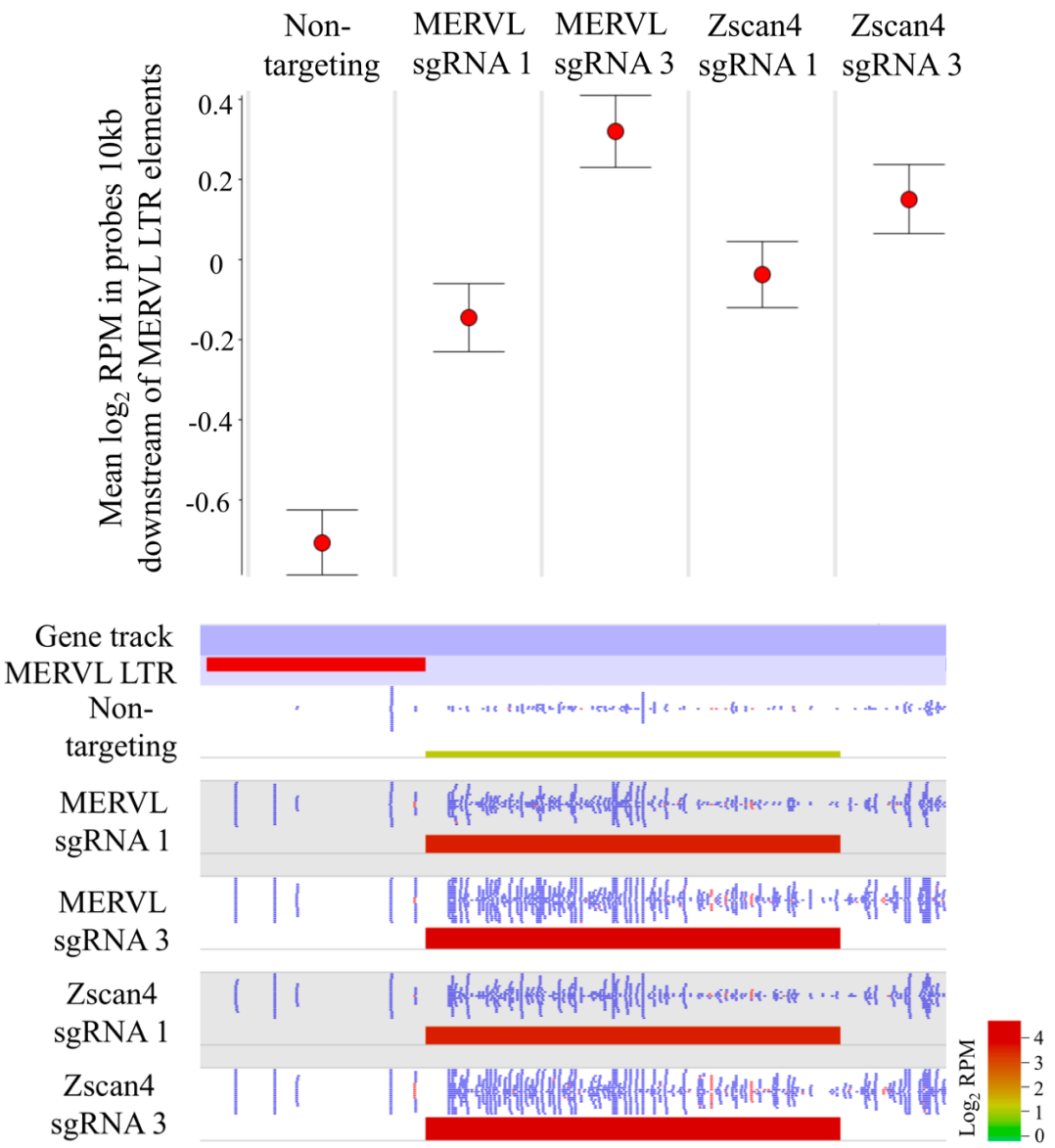



**Figure 3.11– Transcription downstream of MERVL LTR elements after CRISPRa of MERVL and *Zscan4*, analysed by RNA-sequencing**

**Top panel:** Normalised quantification of transcription in 10kb probes downstream of MERVL LTR elements genome-wide, analysed by RNA-sequencing in SAM22 mESCs transduced with a non-targeting sgRNA control, with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3). RNA was harvested 6 days post-sgRNA transduction and puromycin selection. The experiment was performed in triplicate. Data is shown as mean plus and minus standard error.

**Bottom panel:** Representative example of a 10kb probe downstream of a MERVL LTR element in chr7:1082594-10830581, showing raw reads (blue) and quantification in $\log_2$ reads per million (RPM) (green-to-red bars) in SAM22 mESCs transduced with a non-targeting sgRNA control, with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3).

I also studied ZGA-like gene expression in these RNA-sequencing libraries to try to recapitulate the results seen by qPCR; specifically, I analysed the expression of transcripts in the MERVL/ZSCAN4-driven network, described in Eckersley-Maslin et al. 2016, and of transcripts deregulated upon *Dux* induction in mESCs, as described in Hendrickson et al. 2017. DUX overexpression in mESCs has been shown to upregulate a ZGA-like transcriptional signature (Hendrickson et al. 2017; De Iaco et al. 2017). In both cases, most of these ZGA-like transcripts were upregulated upon CRISPRa of MERVL LTRs and *Zscan4* (Figure 3.12). Consistent with my previous results, MERVL sgRNA 3 and Zscan4 sgRNA 1 induced higher expression of these transcripts compared to the other two sgRNAs.

Furthermore, the DEGs upon MERVL and *Zscan4* activation showed a peak of expression in two-cell embryos, when ZGA occurs, corroborating that MERVL and *Zscan4* CRISPRa induces a ZGA-like transcriptional signature (Figure 3.13). I can therefore conclude that the CRISPRa SAM method that I have established in mESCs works effectively to induce endogenous target gene activation and the expected downstream transcriptional changes.

Despite the robustness of the results described above, I further validated CRISPRa target specificity and downstream network activation by comparing it to a more traditional method of gene overexpression, cDNA transfection. Using published RNA-sequencing data of ZSCAN4C transient overexpression (Eckersley-Maslin et al. 2019), I used principal component analysis (PCA) (Pearson 2010) on the whole transcriptome to understand how similar were the responses triggered by CRISPRa and cDNA overexpression of the same target. For transient ZSCAN4C overexpression, E14 mESCs were transfected with ZSCAN4C-eGFP and sorted for $eGFP^{-/+}$ 48 hours later. While principal component 1 (PC1) (48.75% of variation in the data) separated the different datasets by experiment (ZSCAN4C cDNA overexpression and MERVL/*Zscan4* CRISPRa), interestingly, PC2 (16.3% of variation in the data) sorted datasets by control (non-targeting sgRNA and ZSCAN4C-$GFP^{-}$) and experimental conditions (MERVL/*Zscan4* CRISPR-activated cells and ZSCAN4C-$GFP^{+}$ cells) (Figure 3.14). This implies that, after taking experimental conditions into consideration, CRISPRa of endogenous targets promoted a similar transcriptional response to cDNA overexpression.

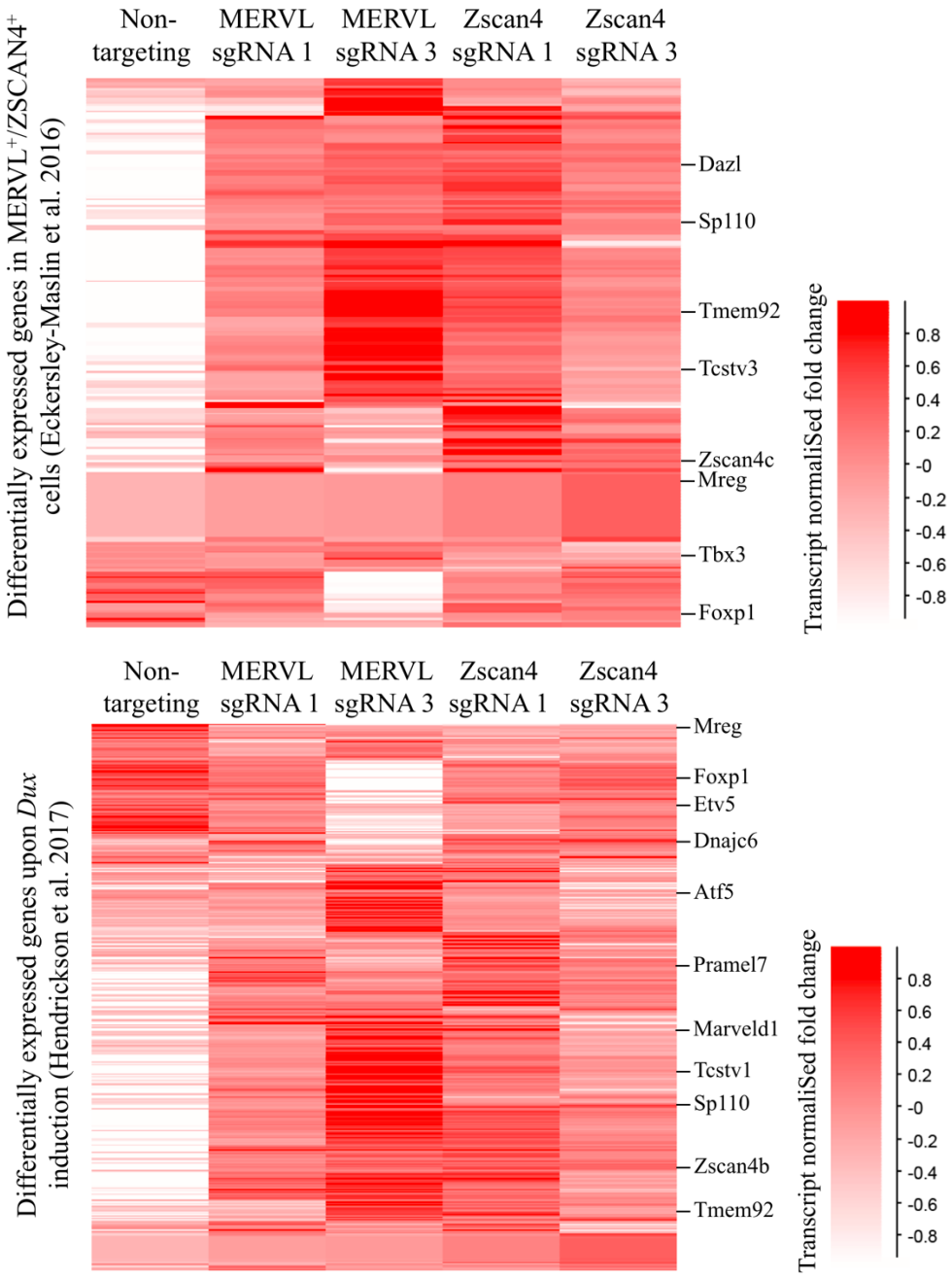



**Figure 3.12– Heatmaps of ZGA-like gene expression after CRISPRa of MERVL and *Zscan4*, analysed by RNA-sequencing**

Heatmaps showing transcript normalised expression, scaled per transcript, of differentially expressed genes (DEGs) in MERVL$^+$/ZSCAN4$^+$ cells (as defined in Eckersley-Maslin et al. 2016) (top panel), and of DEGs upon *Dux* doxycycline induction (analysed from Hendrickson et al. 2017) (bottom panel), in SAM22 mESCs transduced with a non-targeting sgRNA control, with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3), analysed by RNA-sequencing. RNA was harvested 6 days post-sgRNA transduction and puromycin selection. The experiment was performed in triplicate.

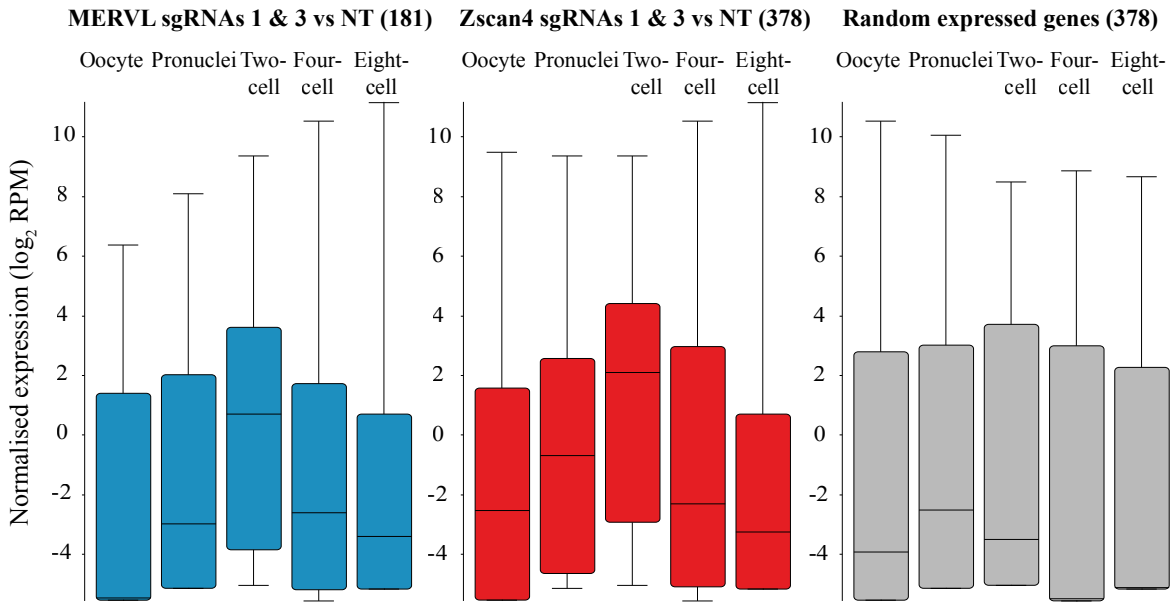



**Figure 3.13– Expression of differentially expressed genes after CRISPRa of MERVL and *Zscan4* in pre-implantation development, analysed by RNA-sequencing**

Box-whisker plots showing normalised expression in $\log_2$ reads per million (RPM) of differentially expressed genes (DEGs) between MERVL sgRNAs and non-targeting sgRNA control (blue), and between Zscan4 sgRNAs and non-targeting control (red), as well as the expression of a random set of expressed genes in the MERVL and *Zscan4* CRISPRa dataset, in mouse oocytes and pre-implantation development (data analysed from Xue et al. 2013). DEGs were determined using DESeq2 (FDR<5%), EdgeR (FDR<5%) and intensity difference filter (FDR<5%), with the high-confidence DEGs defined as the intersection between the three statistical tests. The DEGs by MERVL sgRNA 1 were merged to the DEGs by MERVL sgRNA 3 and the DEGs by Zscan4 sgRNA 1 merged to the DEGs by Zscan4 sgRNA 3 for this analysis. The number of DEGs in each case is depicted in brackets. NT: non-targeting.

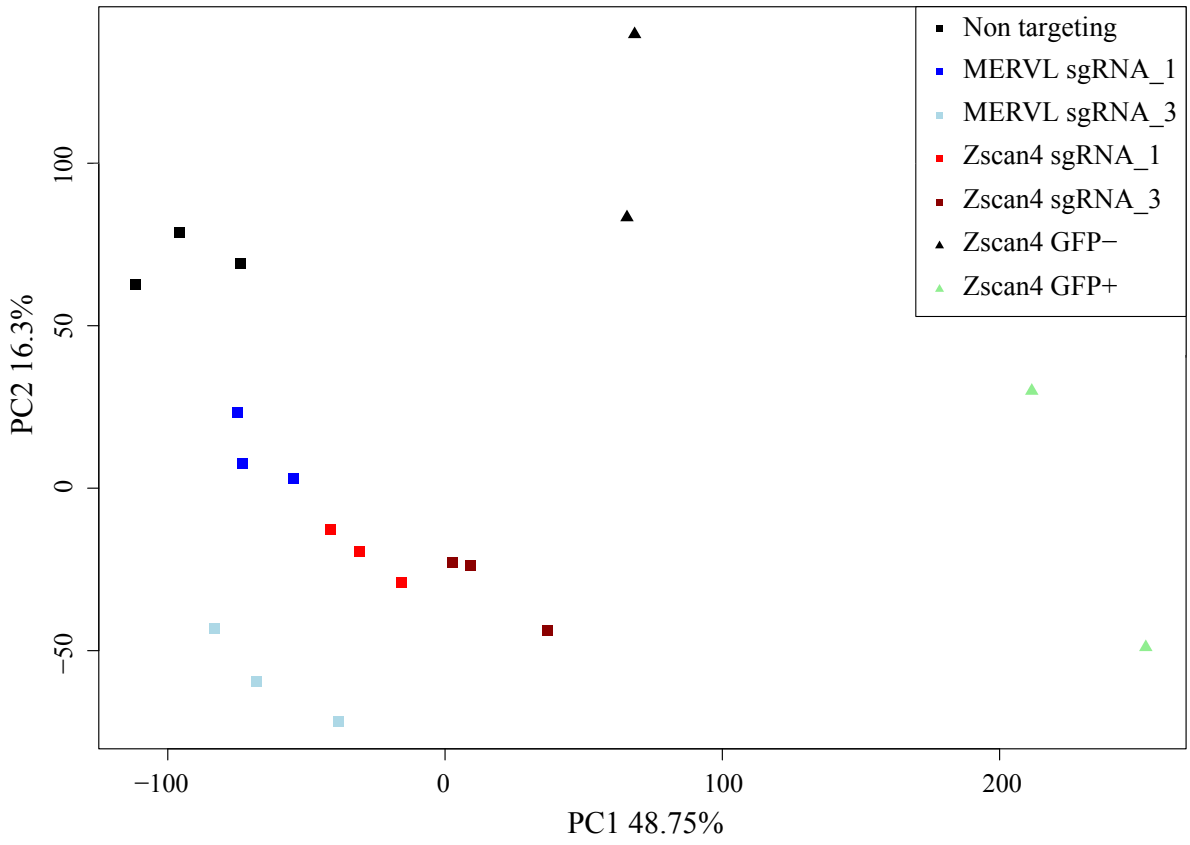



**Figure 3.14– Principal component analysis of MERVL and *Zscan4* CRISPRa and ZSCAN4C cDNA overexpression**

Principal component 1 (PC1) and principal component 2 (PC2) on RNA-sequencing quantified read counts for all transcripts in SAM22 mESCs transduced with a non-targeting sgRNA control (black square), with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 – blue square, and MERVL sgRNA 3 - light blue square) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1- red square, and Zscan4 sgRNA 3- brown square) and E14s mESCs transiently transfected for 48h with ZSCAN4C-eGFP and sorted for $eGFP^{-/+}$ ($ZSCAN4C\text{-}GFP^{-}$-black triangle, $ZSCAN4C\text{-}GFP^{+}$- green triangle). The CRISPRa experiment was performed in triplicate. $ZSCAN4C\text{-}GFP^{-}$ and $ZSCAN4C\text{-}GFP^{+}$ data analysed from Eckersley-Maslin et al. 2019.

The largely different expression levels of the target gene achieved with both methods (Figure 3.15), as well as the differences in timing (6 days for CRISPRa and 2 days for cDNA overexpression), might also account for some of the variation explained by PC1. To further understand the differences between CRISPRa and cDNA overexpression, I looked at the overlap between the DEGs of *Zscan4* CRISPRa, MERVL LTR CRISPRa and ZSCAN4C cDNA overexpression. Consistent with the results presented above, the transcriptome of MERVL LTR CRISPR-activated cells is very similar to *Zscan4* CRISPRa (Figure 3.16).

However, comparison with cDNA overexpression revealed that out of the 494 DEGs between ZSCAN4C-GFP$^{-}$ and ZSCAN4C-GFP$^{+}$, only 32.4% (160/494) were commonly deregulated upon CRISPRa of *Zscan4* and/or MERVL LTRs compared to a non-targeting sgRNA control (Figure 3.16). Intriguingly, 56.4% (102/181) of the DEGs upon CRISPRa of MERVL LTR and 40.7% (154/378) of the DEGs in *Zscan4* CRISPR-activated cells were also deregulated in ZSCAN4C cDNA-overexpressing cells (Figure 3.16). Therefore, a higher proportion of the transcriptional changes triggered by CRISPRa was shared with the changes observed by cDNA overexpression than *vice versa*. This could mean that the artificial levels of mRNA and protein expression resulting from cDNA overexpression leads to transcriptional changes that do not completely reflect the physiological role of the gene of interest.

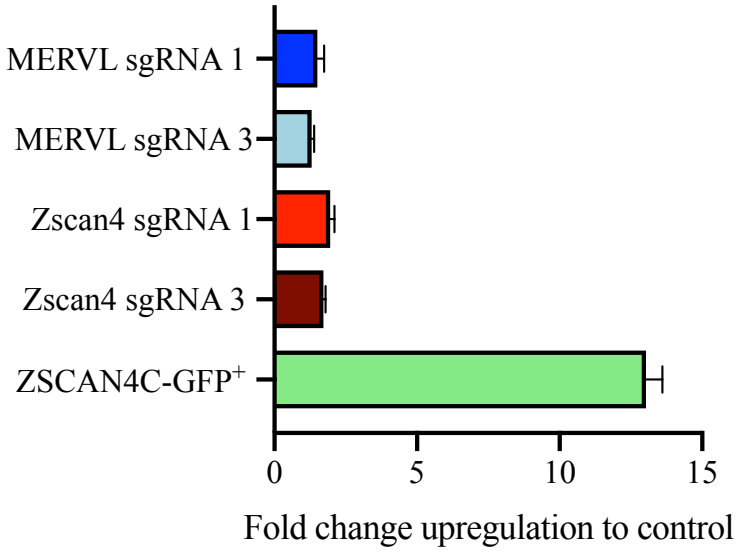



**Figure 3.15– *Zscan4c* normalised expression after CRISPRa or cDNA overexpression**

Fold change expression of *Zscan4c* normalised to corresponding control (non-targeting sgRNA for CRISPRa samples and ZSCAN4-GFP$^{-}$ for ZSCAN4-GFP$^{+}$) in SAM22 mESCs transduced with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA_1 – blue, and MERVL sgRNA_3 - light blue) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA_1- red, and Zscan4 sgRNA_3 - brown) and E14s mESCs transiently transfected for 48h with ZSCAN4-eGFP and sorted for eGFP$^{+}$ (green). The experiment was performed in triplicate and analysed by RNA-sequencing. ZSCAN4C-GFP$^{-}$ and ZSCAN4C-GFP$^{+}$ data analysed from Eckersley-Maslin et al. 2019. Data is shown as mean plus standard deviation.

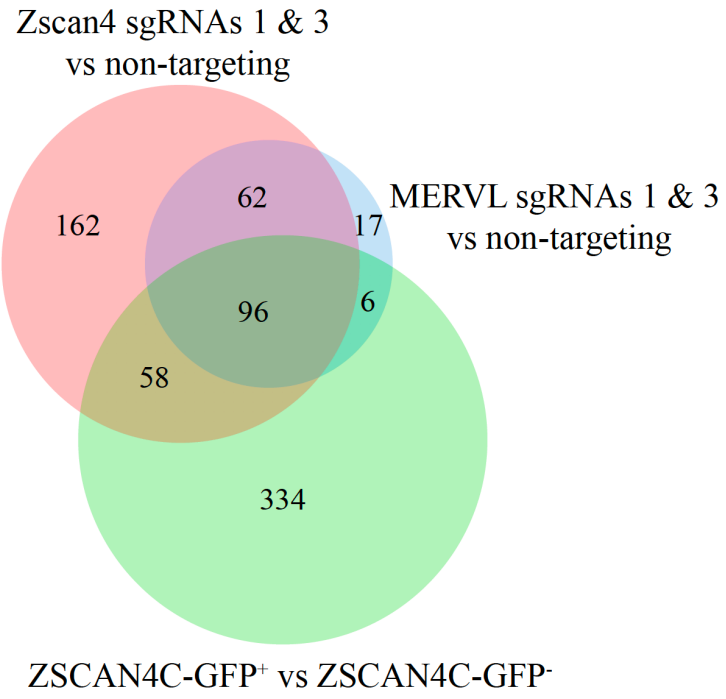



**Figure 3.16– Venn diagram of differentially expressed genes by CRISPRa and cDNA overexpression**

Venn diagram showing the overlap of the differentially expressed genes (DEGs) in 1) ZSCAN4C-GFP$^+$ cells compared to ZSCAN4C-GFP$^-$ (green), 2) MERVL sgRNA 1 and 3 cells compared to non-targeting sgRNA control (blue), and 3) Zscan4 sgRNA 1 and 3 cells compared to non-targeting sgRNA control (pink), as analysed by RNA-sequencing. DEGs were determined using DESeq2 (FDR<5%), EdgeR (FDR<5%) and intensity difference filter (FDR<5%), with the high-confidence DEGs defined as the intersection between the three statistical tests. The DEGs by MERVL sgRNA 1 were merged to the DEGs by MERVL sgRNA 3 and the DEGs by Zscan4 sgRNA 1 merged to the DEGs by Zscan4 sgRNA 3 for this analysis. The experiment was performed in triplicate. ZSCAN4C-GFP$^-$ and ZSCAN4C-GFP$^+$ data analysed from Eckersley-Maslin et al. 2019.

Lastly, I assessed the duration of endogenous target gene activation and downstream changes induced by CRISPRa SAM by performing a time-course experiment in which RNA samples were harvested 6, 9, 11, 13 and 15 days post-sgRNA transduction into SAM22 mESCs. Puromycin selection was maintained throughout the duration of the experiment. Analysis by qPCR showed that MERVL and *Zscan4* transcriptional activation did not significantly change after six days of culture and could be maintained up to 15 days (Figure 3.17). Similarly, the expression of three ZGA markers (*Gm4014*, *Tcstv3* and *Gm4340/Gm2123*), but not of the pluripotency markers *Pou5f1* (also known as *Oct4*) and *Nanog*, was induced to a greater or lesser extent up to 15 days of culture (Figure 3.18).

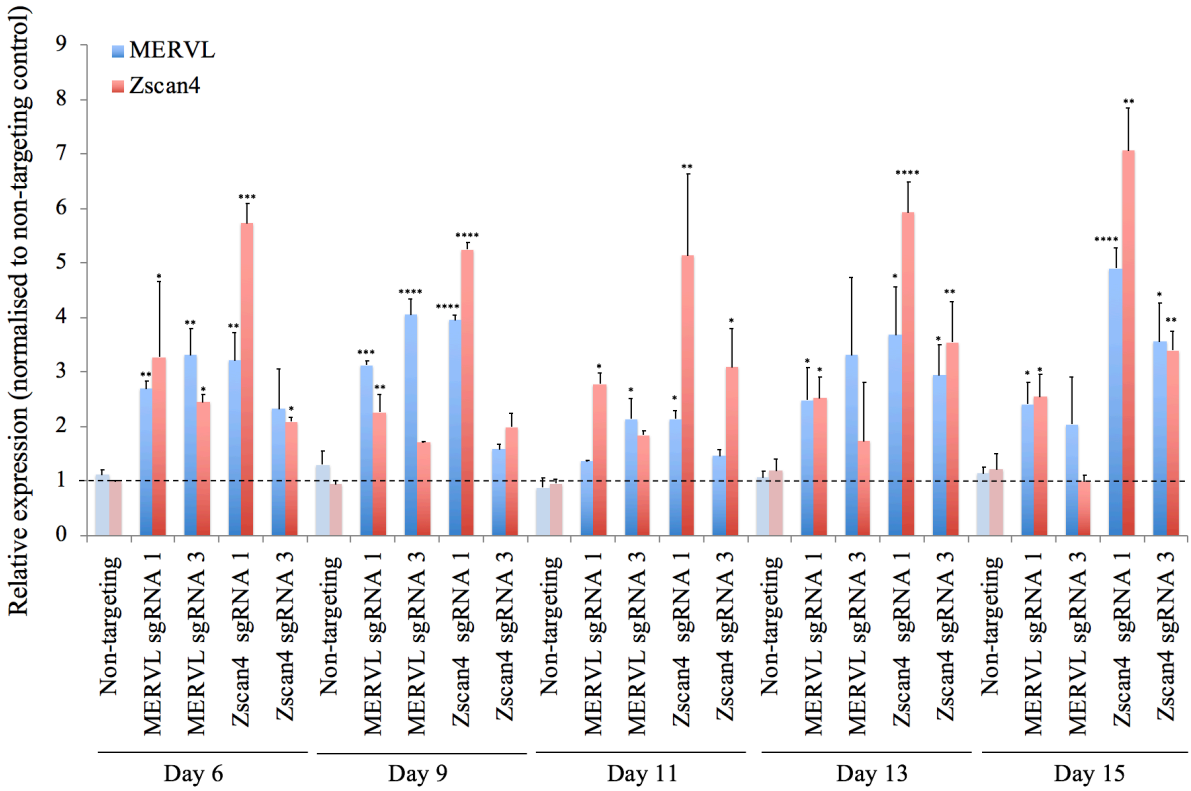



**Figure 3.17– Time-course analysis of MERVL and *Zscan4* expression induced by CRISPRa, analysed by qPCR**

Analysis of MERVL (blue) and *Zscan4* (red) relative RNA levels by qPCR in SAM22 mESCs transduced with a non-targeting sgRNA control (lighter colours), with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) and cultured for 6, 9, 11, 13 or 15 days after sgRNA transduction with puromycin selection. The experiment was performed in triplicate and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to non-targeting controls are reported (homoscedastic two-tailed t-test, ****: p-value < 0.0001, ***: p-value <0.001, **: p-value <0.01, * p-value <0.5, absence of starts: not significant). A dashed line at y = 1 is used to indicate that any condition above it shows transcript upregulation compared to non-targeting control.

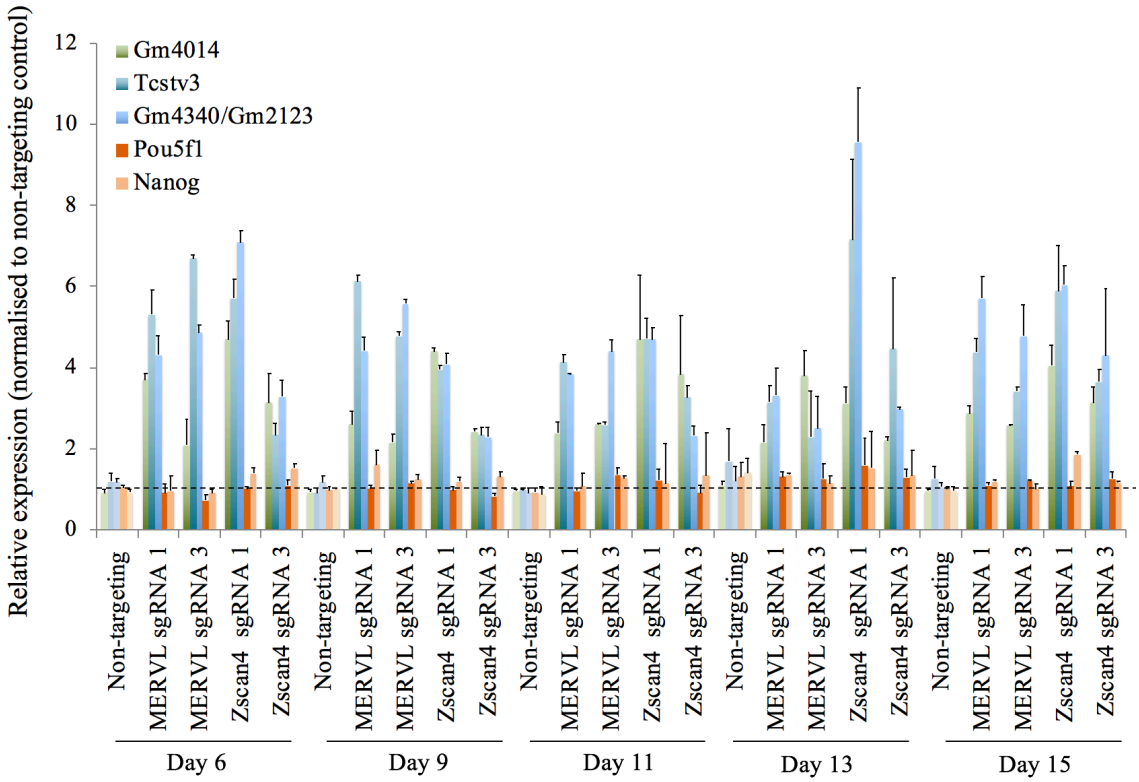



**Figure 3.18– Time course analysis of ZGA markers, *Pou5f1* and *Nanog* expression after CRISPRa of MERVL and *Zscan4*, analysed by qPCR**

Analysis of relative mRNA levels of a panel of ZGA markers (*Gm4014*- green, *Tcstv3*- dark blue, *Gm3230/Gm2123*- light blue), *Pou5f1* (*Oct4*) (dark orange) and *Nanog* (light orange) by qPCR in SAM22 mESCs transduced with a non-targeting sgRNA control (lighter colours), with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) or with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) and cultured for 6, 9, 11, 13 or 15 days after sgRNA transduction with puromycin selection. The experiment was performed in triplicate and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to non-targeting controls were assessed by homoscedastic two-tailed t-test. For figure simplicity, */**/***/**** are not depicted; all transcripts except *Pou5f1* and *Nanog* were significantly upregulated in all experimental conditions compared to the corresponding non-targeting sgRNA control, with a p-value < 0.05. A dashed line at y = 1 is used to indicate that any condition above it shows transcript upregulation compared to non-targeting control.

### 3.2.3 ZGA-like signature detection by single-cell RNA-sequencing

With the aim to combine CRISPRa with a scRNA-seq read-out to ultimately perform a pooled screen, I next assessed whether the target gene activation observed by bulk measurements (qPCR and RNA-sequencing), as well as the downstream changes, could be detected in

scRNA-seq libraries. I used the single-cell 3' solution from 10X Genomics that captures 3' polyadenylated transcripts and allows the transcriptome of thousands of single cells to be sequenced at a low cost per cell (see chapters 1 and 2).

SAM22 mESCs were transduced with each of the same sgRNAs previously described: a non-targeting sgRNA, MERVL sgRNA 1, MERVL sgRNA 3, Zscan4 sgRNA 1 and Zscan4 sgRNA 3. Six days post-transduction and puromycin selection, the cells transduced with MERVL sgRNA 1 and MERVL sgRNA 3 were pooled together, and the cells transduced with Zscan4 sgRNA 1 and Zscan4 sgRNA 3 were pooled together in a second pool. 8,000 single cells from each condition, as well as control E14 and untransduced SAM22 mESCs, were FACS-sorted for live cells, based on the absence of DAPI staining, and 10X scRNA-seq libraries were prepared (see Materials and Methods). Droplet formation in the sample of untransduced SAM22 mESCs failed (data not shown) and was not carried forward. For the other four samples (E14, SAM22 + non-targeting control, SAM22 + MERVL sgRNAs and SAM22 + Zscan4 sgRNAs), between 1,956 and 2,362 single cells were recovered for each library according to detected cellular barcodes, yielding a capture efficiency of 24.5-29.5% (Figure 3.19).

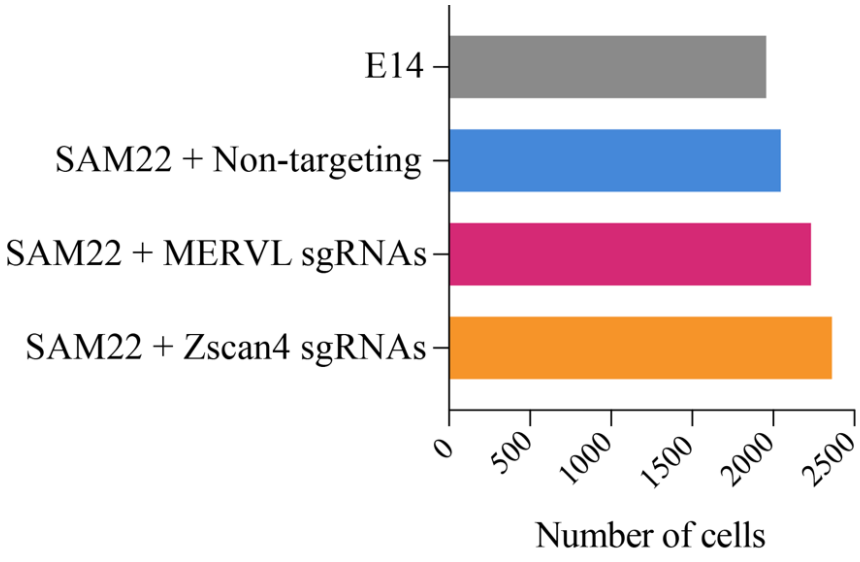



**Figure 3.19– Number of cells captured in 10X Genomics scRNA-seq libraries**

Number of cells captured in 10X Genomics scRNA-seq libraries of E14 mESCs (grey), SAM22 mESCs transduced with a non-targeting sgRNA control (blue), SAM22 mESCs transduced with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) (pink) and SAM22 mESCs transduced with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) (orange).

The distribution of number of UMI reads and number of genes detected for every cell suggested a population of low-quality cells in the four libraries, with less than 15,000 UMI reads and 4,000 detected genes and more than 5% of reads coming from mitochondrial genes (Figure 3.20), likely indicative of sub-optimal sample preparation (see Materials and Methods). These cells were removed for downstream analysis to keep only high-quality input.

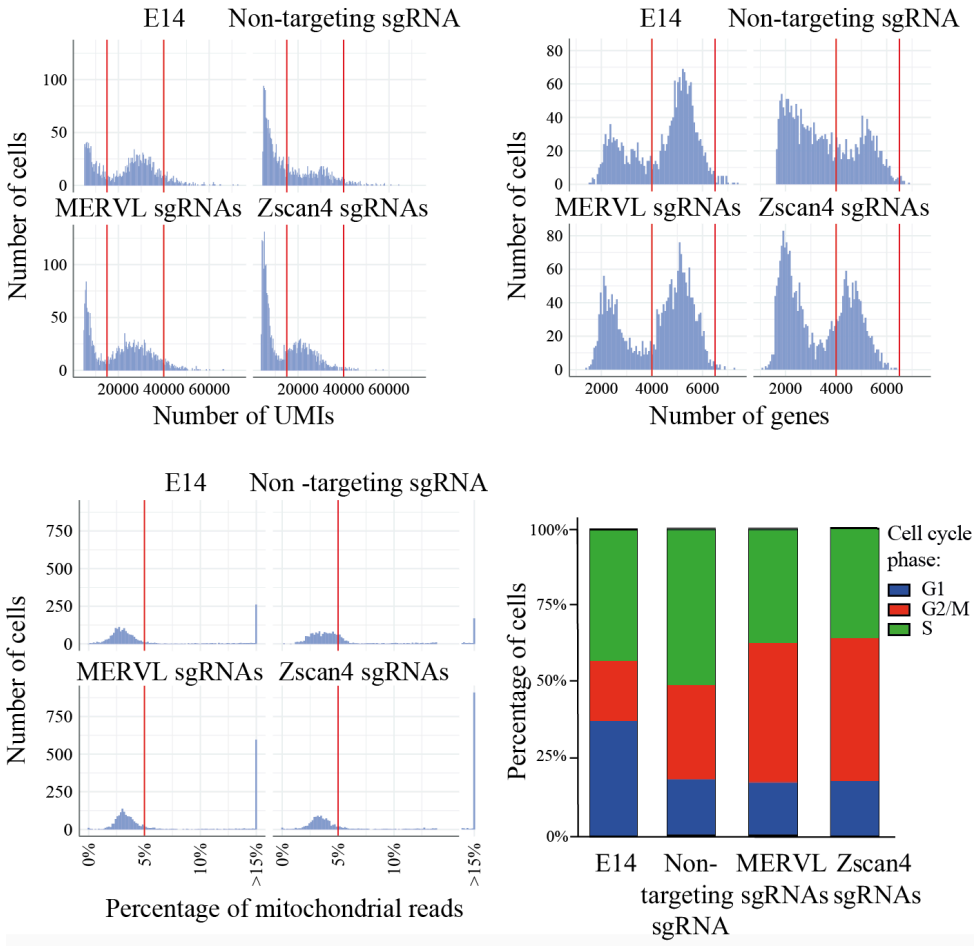



**Figure 3.20– Quality control and filtering of 10X Genomics scRNA-seq libraries**

Histograms showing number of unique molecular identifier (UMIs) (top left panel), number of genes detected (top right panel) and percentage of reads from mitochondrial genes (bottom left panel) in 10X Genomics scRNA-seq libraries of E14 mESCs, SAM22 mESCs transduced with a non-targeting sgRNA control, SAM22 mESCs transduced with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) and SAM22 mESCs transduced with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3). As indicated with red vertical lines, cells with more than 15,000 and less than 40,000 UMIs (top left panel), more than 4,000 and less than 6,500 genes (top right panel) and less than 5% of mitochondrial reads (bottom panel) were retained as high-quality cells. Within the high-quality cells for every sample, the proportion of cells in each phase of the cell cycle was analysed (bottom right panel) as described in Satija et al. 2015.

Within the high-quality cells, cell cycle analysis (as described in Satija et al. 2015) showed that MERVL and *Zscan4* CRISPR-activated cells exist in all stages of the cell cycle, although with a prolonged G2/M phase (Figure 3.20), as previously described for 2C-like cells (Storm et al. 2014, Eckersley-Maslin et al. 2016). Next, I assessed target gene activation in every cell by analysing MERVL and *Zscan4* cluster expression. These key markers of ZGA were expressed in 1.93% and 1.23% of untransduced E14 mESCs respectively, consistent with previous studies (Macfarlan et al. 2012; Zalzman et al. 2010) (Figures 3.21, 3.22). Upon transduction with a non-targeting sgRNA control, the population of cells expressing MERVL and *Zscan4* increased to 3.49% and 18.63%, respectively, indicating that lentiviral transduction of mESCs slightly increased the expression of these markers (Figures 3.21, 3.22). Notably, and consistent with my previous results, 10.43% of cells transduced with MERVL LTR sgRNAs and 52.8% of cells transduced with Zscan4 sgRNAs activated their respective targets (Figures 3.21, 3.22). Strikingly, MERVL CRISPRa induced *Zscan4* expression and *vice versa*, consistent with my previous qPCR and bulk RNA-sequencing results (Figures 3.5, 3.6, 3.7, 3.9, 3.10, 3.11, 3.17).

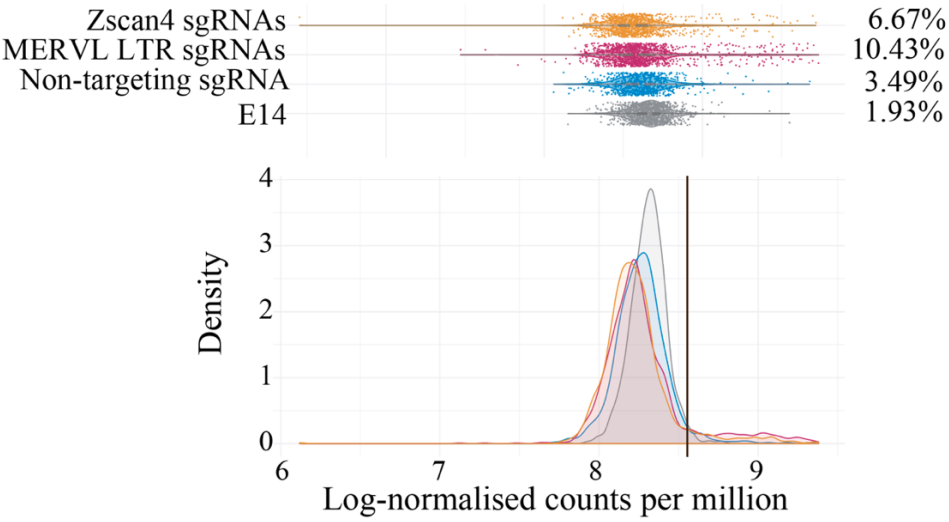



**Figure 3.21– MERVL expression after CRISPRa of MERVL and *Zscan4*, analysed by 10X Genomics scRNA-seq**

Dot plot (top panel) and density representation (bottom panel) of MERVL normalised expression in each single cell (in $\log_e$(number of UMIs for MERVL / sum (number of total UMI in a cell) *10,000 + 1)) in E14 mESCs (grey), SAM22 mESCs transduced with a non-targeting sgRNA control (blue), SAM22 mESCs transduced with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) (pink) and SAM22 mESCs transduced with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) (orange), analysed by 10X Genomics scRNA-seq. Reported percentages of cells expressing MERVL are calculated as the proportion of cells above the vertical black line out the total number of cells in each sample.

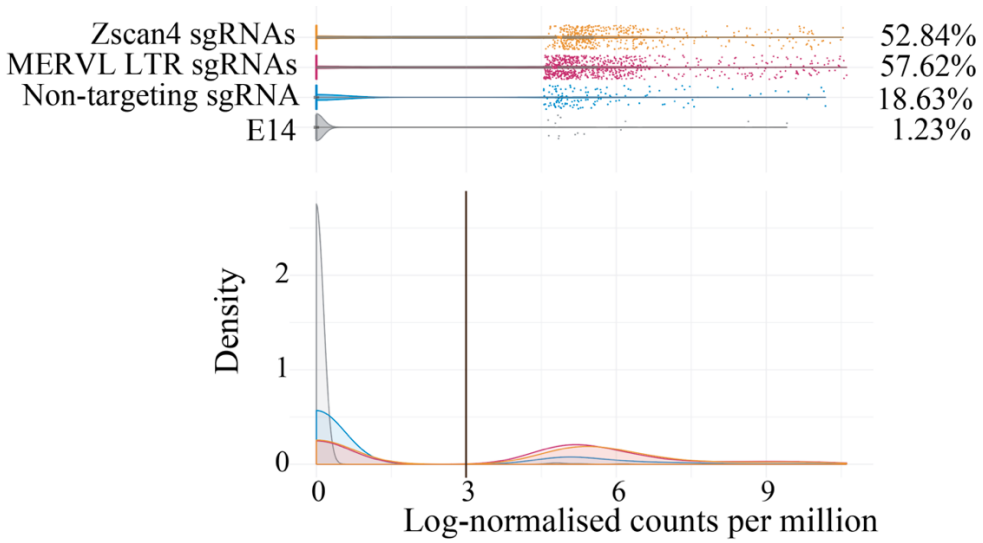



**Figure 3.22– *Zscan4* expression after CRISPRa of MERVL and *Zscan4*, analysed by 10X Genomics scRNA-seq**

Dot plot (top panel) and density representation (bottom panel) of Zscan4 (all genes) normalised expression in each single cell (in loge(number of UMIs for *Zscan4b/c/d/e/f* / sum (number of total UMIs in a cell) *10,000 + 1)) in E14 mESCs (grey), SAM22 mESCs transduced with a non-targeting sgRNA control (blue), SAM22 mESCs transduced with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) (pink) and SAM22 mESCs transduced with sgRNAs targeted to the promoters of the Zscan4 cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) (orange), analysed by 10X Genomics scRNA-seq. Reported percentages of cells expressing *Zscan4* are calculated as the proportion of cells above the vertical black line out the total number of cells in each sample.

This scRNA-seq experiment also revealed that not all cells selected for sgRNA expression activated their targets, consistent with the immunofluorescence results for ZSCAN4 (Figure 3.7). Reassuringly, nearly all cells in all conditions had high expression of the pluripotency marker *Pou5f1* (also known as *Oct4*) (Figure 3.23), consistent with previous reports showing that 2C-like cells lack OCT4 protein but the mRNA is not downregulated (Macfarlan et al. 2012; Ishiuchi et al. 2015; Eckersley-Maslin et al. 2016; Choi et al. 2017; Rodriguez-Terrones et al. 2018).

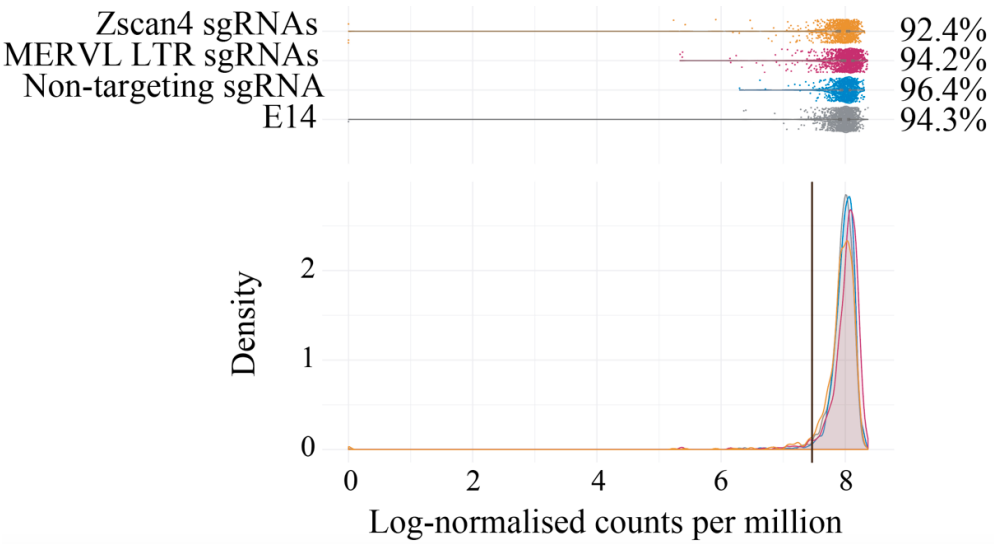



**Figure 3.23– *Pou5f1* expression after CRISPRa of MERVL and *Zscan4*, analysed by 10X Genomics scRNA-seq**

Dot plot (top panel) and density representation (bottom panel) of *Pou5f1* (*Oct4*) normalised expression in each single cell (in loge(number of UMIs for *Pou5f1* / sum (number of total UMsI in a cell) *10,000 + 1)) in E14 mESCs (grey), SAM22 mESCs transduced with a non-targeting sgRNA control (blue), SAM22 mESCs transduced with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) (pink) and SAM22 mESCs transduced with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) (orange), analysed by 10X Genomics scRNA-seq. Reported percentages of cells expressing *Pou5f1* are calculated as the proportion of cells above the vertical black line out the total number of cells in each sample.

Similar to bulk RNA-sequencing analysis, I then looked at the expression of the MERVL/ZSCAN4-driven network described in Eckersley-Maslin et al. 2016 (Figure 3.24) and of the transcripts induced upon DUX overexpression (Hendrickson et al. 2017) (Figure 3.25). Between 2% and 2.46% of untransduced E14 mESCs expressed these ZGA transcripts, consistent with previous studies (Macfarlan et al. 2012; Zalzman et al. 2010; Eckersley-Maslin et al. 2016; Hendrickson et al. 2017; De Iaco et al. 2017; Eckersley-Maslin et al. 2019). This proportion was very slightly increased upon transduction of SAM22 mESCs with a non-targeting sgRNA control (Figures 3.24, 3.25), an observation that has also been made in the Reik laboratory when mESCs are transiently transfected with GFP (Eckersley-Maslin, M., personal communication). Upon CRISPRa of MERVL LTRs or *Zscan4*, the proportion of cells expressing these ZGA-like transcripts increased to 6%-10% (Figure 3.24, 3.25), regardless of

the target gene activation state in each individual cell. These results confirm that a ZGA-like signature can be detected by scRNA-seq and validates my approach of using CRISPRa in mESCs combined with scRNA-seq as a system to screen for ZGA regulators.

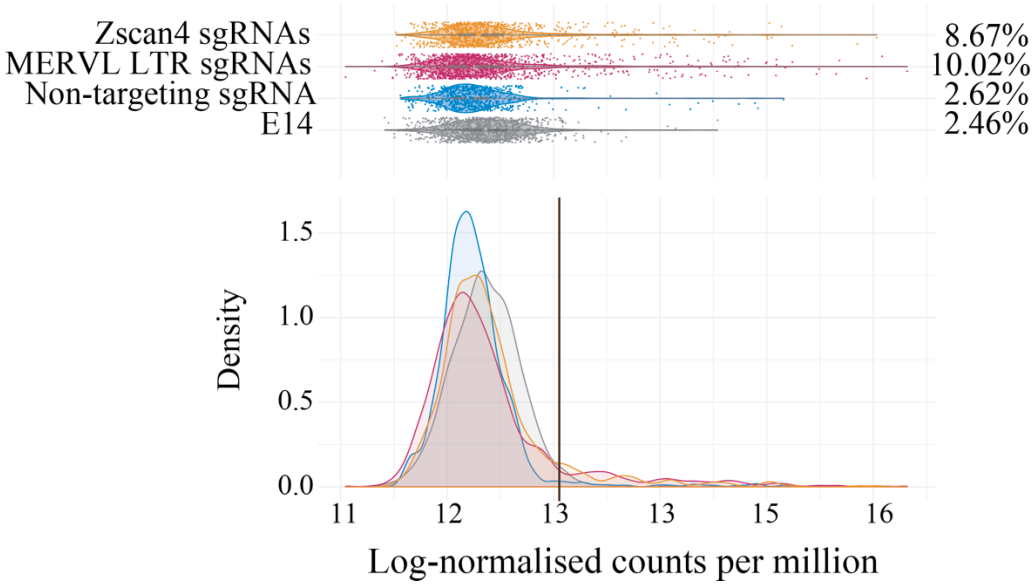



**Figure 3.24– MERVL/ZSCAN4 network expression after CRISPRa of MERVL and *Zscan4*, analysed by 10X Genomics scRNA-seq**

Dot plot (top panel) and density representation (bottom panel) of the expression of differentially expressed genes (DEGs) in MERVL$^{+}$/ZSCAN4$^{+}$ cells (as defined in Eckersley-Maslin et al. 2016) in each single cell (in $\log_e$(number of UMIs / sum (number of total UMsI in a cell) *10,000 + 1)) in E14 mESCs (grey), SAM22 mESCs transduced with a non-targeting sgRNA control (blue), SAM22 mESCs transduced with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) (pink) and SAM22 mESCs transduced with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) (orange), analysed by 10X Genomics scRNA-seq. Reported percentages of cells expressing the indicated transcripts are calculated as the proportion of cells above the vertical black line out the total number of cells in each sample.

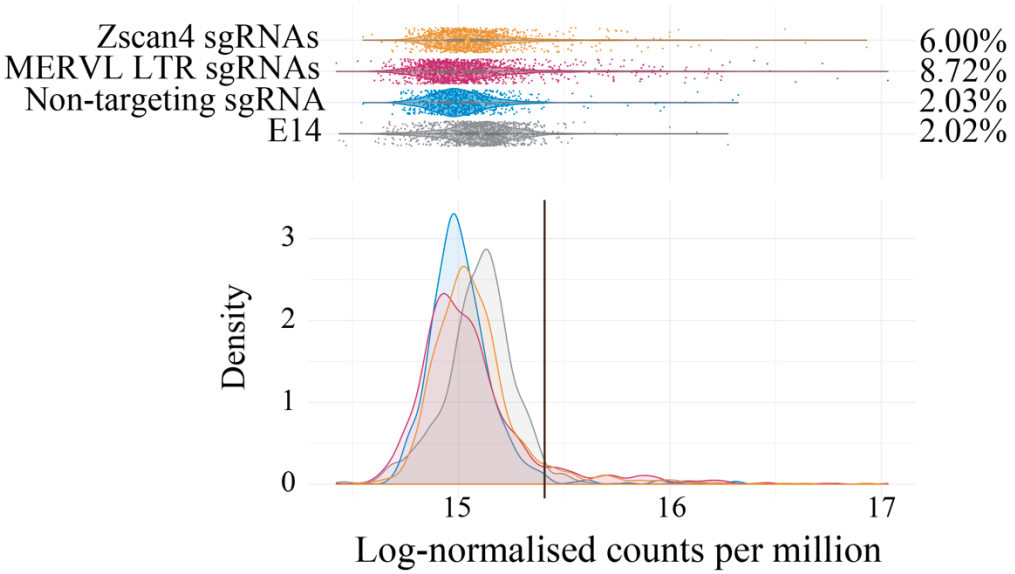



**Figure 3.25– Expression of DUX-regulated genes after CRISPRa of MERVL and *Zscan4*, analysed by 10X Genomics scRNA-seq**

Dot plot (top panel) and density representation (bottom panel) of the expression of differentially expressed genes (DEGs) upon *Dux* doxycycline induction (analysed from Hendrickson et al. 2017) in each single cell (in $\log_e$(number of UMIs / sum (number of total UMsI in a cell) *10,000 + 1)) in E14 mESCs (grey), SAM22 mESCs transduced with a non-targeting sgRNA control (blue), SAM22 mESCs transduced with sgRNAs targeted to MERVL LTR elements (MERVL sgRNA 1 and MERVL sgRNA 3) (pink) and SAM22 mESCs transduced with sgRNAs targeted to the promoters of the *Zscan4* cluster (Zscan4 sgRNA 1 and Zscan4 sgRNA 3) (orange), analysed by 10X Genomics scRNA-seq. Reported percentages of cells expressing the indicated transcripts are calculated as the proportion of cells above the vertical black line out the total number of cells in each sample.

Contrary to what I observed in bulk RNA-sequencing analysis (Figure 3.13), the DEGs upon MERVL LTR and *Zscan4* CRISPRa, as analysed by scRNA-seq, did not show a clear peak of expression in two-cell embryos (Figure 3.26). This is probably due to the less stringent differential gene expression analysis applied to scRNA-seq data (only EdgeR FDR<5% *versus* DESeq2 FDR<5%, EdgeR FDR<5% and intensity difference FDR<5% in bulk RNA-sequencing data), since filtering for the top 100 differentially upregulated genes showed a similar pattern of expression in pre-implantation embryos to that observed by bulk RNA-sequencing (Figure 3.13, 3.26).

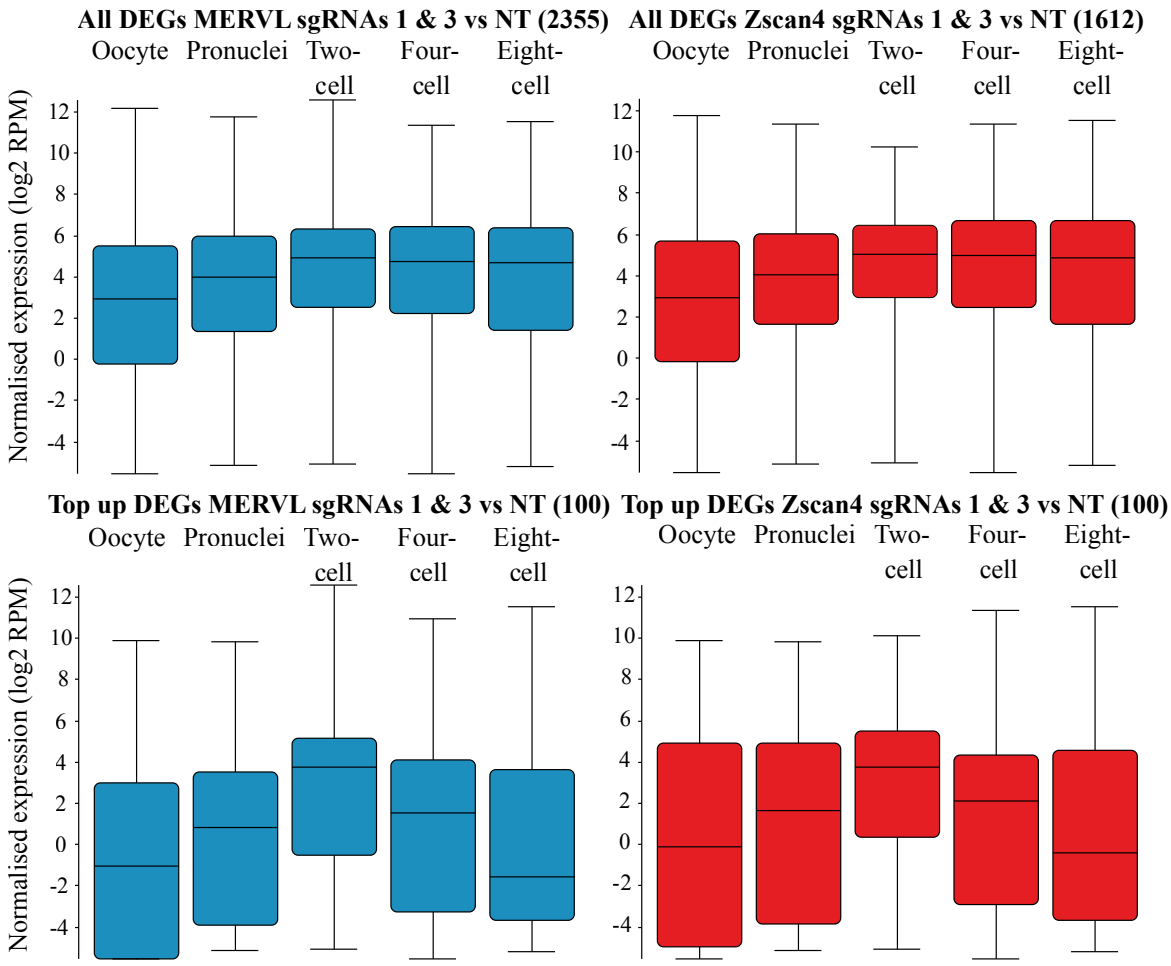



**Figure 3.26– Expression of differentially expressed genes after CRISPRa of MERVL and *Zscan4* in pre-implantation development, analysed by scRNA-seq**

Box-whisker plots showing normalised expression in $\log_2$ reads per million (RPM) of differentially expressed genes (DEGs) between MERVL sgRNAs and non-targeting sgRNA control (blue, top), and between Zscan4 sgRNAs and non-targeting sgRNA control (red, top), as well as the top 100 upregulated (up) DEGs between MERVL sgRNAs and non-targeting sgRNA control (blue, bottom), and the top 100 upregulated DEGs between Zscan4 sgRNAs and non-targeting sgRNA control (red, bottom), in mouse oocytes and pre-implantation development (data analysed from Xue et al. 2013). DEGs were determined using EdgeR (FDR<5%). The number of analysed genes in each case is depicted in brackets. NT: non-targeting

Notwithstanding the power of combining CRISPR with scRNA-seq, the maximum number of candidate genes screened in previous studies is 100 (Jaitin et al. 2016; Dixit et al. 2016; Adamson et al. 2016; Datlinger et al. 2017; S. Xie et al. 2017), which means that the high sequencing costs are an important limitation. Therefore, determining the minimum number of cells required to be sequenced to be able to detect a specific transcriptional signature is crucial to define the number of genes that can be screened. These estimations are highly dependent on the biological context under investigation and on the prominence of the transcriptional signature triggered. This pilot scRNA-seq experiment allowed such power estimation to be made in order to determine the minimum number of cells required to be sequenced per sgRNA in the pooled screen to detect a ZGA-like signature in a potential positive hit. I applied a qualitative two-tailed Fisher's exact test and calculated the sample size necessary to detect a ZGA-like signature with an adjusted p-value < 0.0005 and a power of 0.8, considering that the percentage of cells that expressed ZGA-like transcripts was 2.62% in cells transduced with a non-targeting sgRNA control and 10.02% in cells transduced with a positive inductor (e.g. MERVL LTR sgRNAs) (Figure 3.24). The test returned a sample size of 367 cells required to be screened per sgRNA to detect a positive hit with an actual adjusted p-value < 0.00032.

### 3.2.4 Capturing sgRNA sequence information by single-cell RNA-sequencing

Having optimised CRISPRa with scRNA-seq read-out, I next envisaged a strategy to capture and sequence the sgRNA transduced in each cell by scRNA-seq. This would allow a multiplexed screen to be performed where a pooled sgRNA library would be transduced into SAM22 mESCs and, after antibiotic selection and expansion, scRNA-seq could be used to infer the sgRNA transduced in every cell, consequently the target gene activated and, ultimately, the transcriptional changes triggered by the activation or upregulation of that gene.

Most droplet-based scRNA-seq technologies, including 10X Genomics 3'sequencing, rely on transcripts being polyadenylated at their 3' end for capture with oligo(d)T adaptors (X. Zhang et al. 2019), which ensures capturing mRNAs but makes these methods unsuitable to study non-coding transcription (see section 1.3). sgRNAs are often transcribed from a U6 promoter which is bound by polymerase III (Pol III), leading to a non- 3' polyadenylated transcript (H.

Ma et al. 2014). Recent studies have described ways to overcome this limitation and capture sgRNA sequences in scRNA-seq libraries (Jaitin et al. 2016; Dixit et al. 2016; Adamson et al. 2016; Datlinger et al. 2017; Replogle et al. 2018). CROP-seq allows simultaneous analysis of the transcriptome and the sgRNA transduced into each cell by incorporating the sgRNA transcript within a polymerase II (Pol II) transcript (Datlinger et al. 2017) (Figure 1.18). The strategy consists on designing a lentiviral vector that contains the U6 Pol III promoter and sgRNA cassette within the 3'LTR, at the same position of the deletion that makes lentivirus self-inactivating and just before the 3' LTR polyadenylation signal. Upstream of the 3' LTR, contained within the lentiviral integration site, a Pol II promoter would drive expression of an antibiotic-resistance cassette. Upon lentiviral reverse transcription and integration into the host genome, the U6-sgRNA cassette would be copied to the 5' LTR. Functional sgRNAs would be expressed from the U6 promoter but also, the U6-sgRNA cassette would be contained within the Pol II polyadenylated transcript, making it detectable by scRNA-seq approaches (Figure 1.18).

In order to adapt the CROP-seq strategy to the CRISPRa SAM system, a lentiviral vector was re-synthesize from the CROP-seq backbone with the following modifications: 1) the sgRNA scaffold sequence contains two MS2 loops to allow recruitment of MS2-p65-HSF1 in SAM22 mESCs, same as in the sgRNA(MS2)_puro backbone used in sections 3.2.2 and 3.2.3; and 2) a fluorescent mCherry marker was included downstream of the Pol II EF1α promoter and linked through the self-cleaving peptide T2A to the puromycin resistant cassette to allow assessment of the MOI by flow cytometry. I refer to this construct as CROP-sgRNA-MS2 (Figure 3.27, Appendix A).

To validate this construct for its use in scRNA-seq experiments, I transduced SAM22 mESCs with a non-targeting sgRNA control cloned in either the lentiviral vector sgRNA(MS2)_puro or in the lentiviral construct CROP-sgRNA-MS2. I first had to optimise the lentiviral transduction protocol for sgRNAs cloned into the CROP-sgRNA-MS2 construct, given its lower titer, compared to the sgRNA(MS2)_puro vector (see Materials and Methods, Datlinger et al. 2017). Part of the optimisation included culturing cells with puromycin selection up to 9 days post-transduction before harvesting for analysis, a timepoint where CRISPRa worked effectively with the sgRNA(MS2)_puro backbone (Figure 3.17 and 3.18). After puromycin

selection, I harvested RNA and synthesised cDNA using oligo(d)T adaptors so that only polyadenylated RNA molecules would go through the first strand synthesis of cDNA, a method that recapitulates cDNA synthesis in 10X Genomics 3' scRNA-seq library preparation. To test whether the sgRNA transcribed from the CROP-sgRNA-MS2 construct was polyadenylated, I performed a qPCR on the cDNA synthesised with oligo(d)T adaptors using primers that bound the scaffold sequence of the sgRNA. As expected, the qPCR amplified the sgRNA produced from the CROP-sgRNA-MS2 lentiviral construct but not the sgRNA transcribed from the sgRNA(MS2)_puro construct (Figure 3.28), validating my adaptation of the CROP-seq strategy to the SAM system as a method to perform CRISPRa with scRNA-seq read-out.

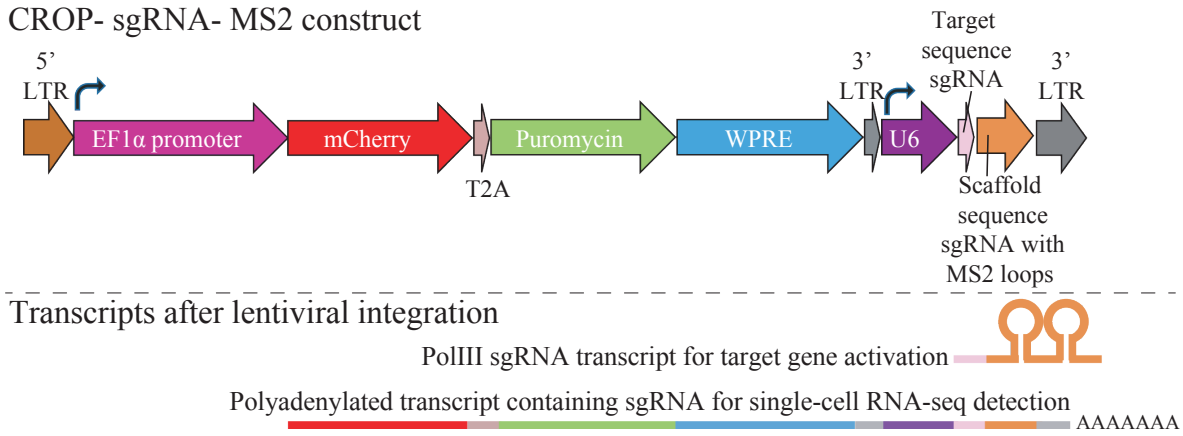



**Figure 3.27– Schematic representation of CROP-sgRNA-MS2 construct**

The integration site of the CROP-sgRNA-MS2 construct contains a Pol II EF1α promoter driving the expression of mCherry and puromycin, linked through the self-cleaving peptide T2A, and of the Woodchuck hepatitis virus posttranscriptional regulatory element (WPRE). Contained within the 3' LTR are the U6 promoter, the target sgRNA sequence and the scaffold sgRNA sequence with two MS2 loops. Upon reverse transcription and lentiviral integration into SAM22 mESCs, two transcripts are produced from this construct: a functional sgRNA Pol III transcript that drives target gene activation and a polyadenylated Pol II transcript containing mCherry, T2A, puromycin, WPRE, U6 promoter and the sgRNA sequence.

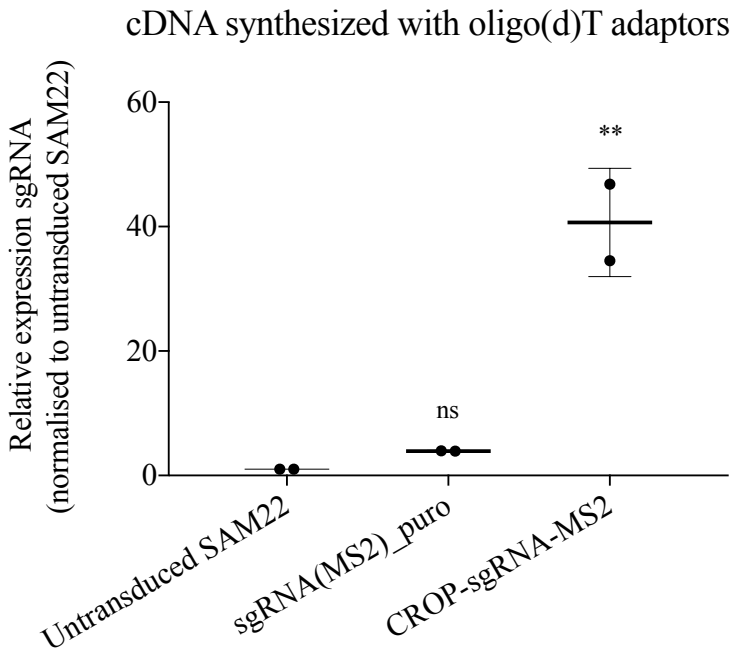



**Figure 3.28– sgRNA capture with oligo(d)T adaptors**

Analysis of polyadenylated sgRNA expression by qPCR in SAM22 mESCs untransduced, transduced with a non-targeting sgRNA cloned in the sgRNA(MS2)_puro lentiviral construct or with a non-targeting sgRNA cloned in the CROP-sgRNA-MS2 lentiviral construct. SAM22 mESCs were cultured for 9 days with puromycin selection after sgRNA transduction. cDNA was synthesised from DNaseI-treated RNA using oligo(d)T primers. Primers used for qPCR analysis bound to the scaffold sequence of the sgRNA. The experiment was performed twice and the qPCR run in two technical replicates. Data is shown as mean plus and minus standard deviation with individual data points shown. Statistically significant differences to untransduced SAM22 mESCs were assessed by homoscedastic two-tailed t-test. ns: non-significant, **: p-value <0.01.

Although the CROP-seq vector had been shown to work effectively for CRISPR KO (Datlinger et al. 2017), I tested whether promoter-targeted sgRNAs cloned into the CROP-sgRNA-MS2 construct could induce effective target gene activation. To that end, I cloned Zscan4 sgRNA 1 and MERVL sgRNA 1 into this new construct as positive controls, since they worked before for inducing target gene activation in the sgRNA(MS2)_puro backbone. I also tested sgRNAs for *Dppa4*, *Kdm5b*, *Dux* and *Dppa3*, as they have been implicated in ZGA regulation (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Dahl et al. 2016; B. Zhang et al. 2016; Hendrickson et al. 2017; De Iaco et al. 2017; Whiddon et al. 2017; Arakawa et al. 2015; Y. Huang et al. 2017; Payer et al. 2003; Bortvin et al. 2004; Y. Li et al. 2018) and, therefore, were likely to be included as controls in the screen. For all the sgRNAs tested, except for Dux sgRNA

2, I observed at least a 2-fold increase in the expression of the target gene compared to a non-targeting sgRNA control by qPCR (Figure 3.29), confirming that insertion of the U6-sgRNA cassette within the 3' LTR does not interfere with sgRNA function.

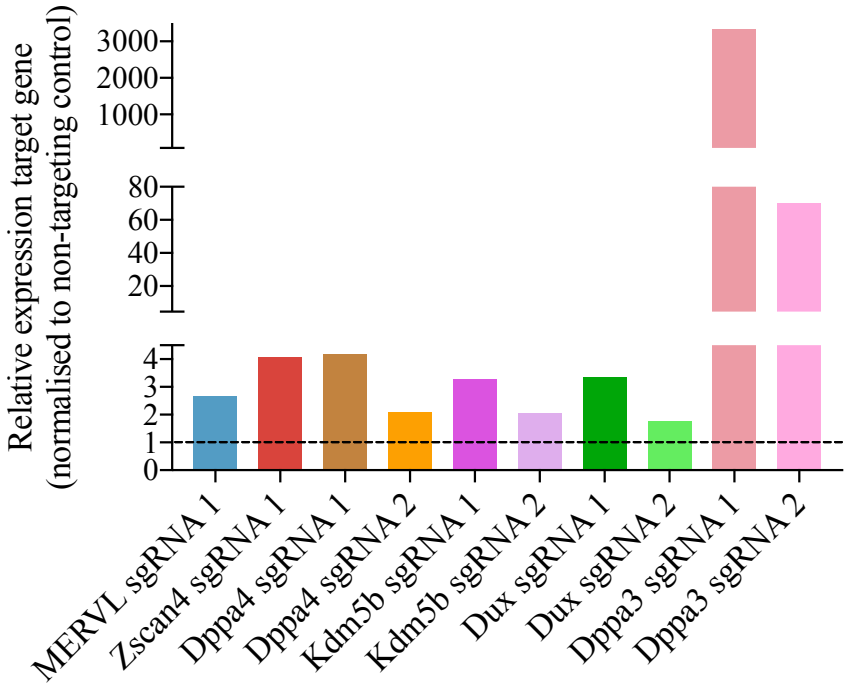



**Figure 3.29– Expression of target genes after CRISPRa with CROP-sgRNA-MS2 construct, analysed by qPCR**

Analysis of relative RNA levels of target genes (MERVL, *Zscan4*, *Dppa4*, *Kdm5b*, *Dux* and *Dppa3*) by qPCR in SAM22 mESCs after CRISPRa. Expression was normalised to average expression in two non-targeting sgRNA controls. Bars show relative expression of the gene targeted by the sgRNA indicated in the x axis. SAM22 mESCs were transduced with two non-targeting sgRNA controls (non-targeting 1 and non-targeting 2), with MERVL LTR sgRNA 1 (blue), with Zscan4 sgRNA 1 (red), with Dppa4 sgRNA 1 (brown), with Dppa4 sgRNA 2 (orange), with Kdm5b sgRNA 1 (dark purple), with Kdm5b sgRNA 2 (light purple), with Dux sgRNA 1 (dark green), with Dux sgRNA 2 (light green), with Dppa3 sgRNA 1 (dark pink) or with Dppa3 sgRNA 2 (light pink), all cloned into the CROP-sgRNA-MS2 construct. RNA was harvested 9 days post-sgRNA transduction and puromycin selection. The experiment was performed once and the qPCR run in two technical replicates. A dashed line at y = 1 is used to indicate that any sgRNA above it showed target transcript upregulation compared to non-targeting control.

This experiment also showed that the effectiveness of CRISPRa varies significantly from one target gene to another and does not necessarily correlate with the basal expression of the target gene, since MERVL, *Zscan4* and *Dux* are expressed in 1-5% of mESCs (Macfarlan et al. 2012; Zalzman et al. 2010; Eckersley-Maslin et al. 2016; Hendrickson et al. 2017; De Iaco et al. 2017)

(Figures 3.7, 3.21 and 3.22 for *Zscan4* and MERVL), *Dppa3* is expressed in 20-30% of mESCs (K. Hayashi et al. 2008) and *Dppa4* and *Kdm5b* are highly expressed to similar levels of the pluripotency marker *Oct4* (Bortvin et al. 2003; L. Xie et al. 2011).

Importantly, I showed previously that RNA upregulation by CRISPRa translated into protein overexpression (Figure 3.6 and 3.7). CRISPRa mediated by sgRNAs expressed from the CROP-sgRNA-MS2 construct also translated into protein overexpression, as shown by immunofluorescence of DPPA3 in cells transduced with the corresponding targeting sgRNA (Figure 3.30). Similar to what was observed for ZSCAN4 protein expression, *Dppa3* CRISPRa did not induce protein upregulation in all the cells selected for the expression of the sgRNA. The reasons why sgRNA expression might induce target upregulation in some cells but not in others are discussed in section 3.3 of this chapter and further in chapter 4 of this dissertation.

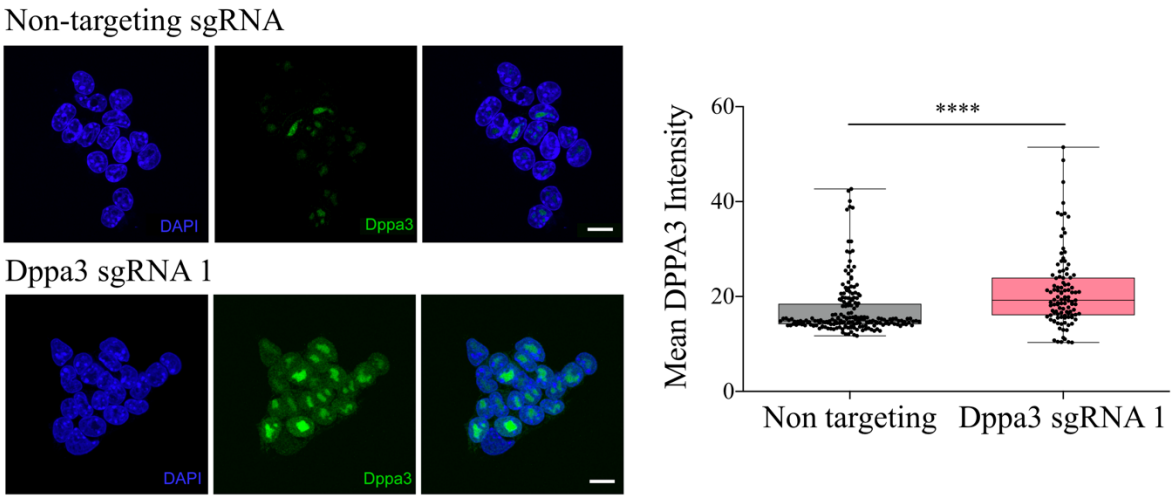



**Figure 3.30– Immunofluorescence of DPPA3 after CRISPRa with CROP-sgRNA-MS2 construct**

**Left panel:** Representative images of SAM22 mESCs transduced with a non-targeting sgRNA control, or with Dppa3 sgRNA 1, both cloned into the CROP-sgRNA-MS2 construct, and stained for DPPA3 protein (green) and DNA stained with DAPI (blue). The cells were seeded on coverslips 15 days after sgRNA transduction and puromycin selection and cultured for an extra day before fixation. Images show a single Z-plane. Scale bar = 10 μm.

**Right panel:** Box-whisker plots showing quantification of DPPA3 protein as mean fluorescence intensity in SAM22 mESCs transduced with a non-targeting sgRNA control (grey), or with Dppa3 sgRNA 1 (pink). A total of 190 cells were quantified in the non-targeting sgRNA control condition and 107 cells in the Dppa3 sgRNA 1 condition. Differences between the two groups were statistically significant (Mann-Whitney test: **** p-value <0.0001).

Having validated the CROP-sgRNA-MS2 construct for sgRNA capture with oligo(d)T adaptors and for effective target gene activation, I performed a pilot scRNA-seq experiment to test sgRNA detection. A major limitation of the CROP-seq strategy is that sgRNA detection in a single cell is highly dependent on cell sequencing coverage and, therefore, many cells with an undetected sgRNA have to be discarded for downstream analysis (Datlinger et al. 2017). In order to overcome this limitation, Hills and colleagues developed a strategy to increase the proportion of cells assigned to a sgRNA in 10X Genomics 3' scRNA-seq libraries to 94% (Hill et al. 2018). The method consists on performing a PCR enrichment for barcoded cDNA fragments that contain a sgRNA, using three rounds of amplification (Figure 3.31): in the first round, the forward primer binds the U6 promoter which is upstream of the sgRNA, and the reverse primer binds half of the Read 1 Nextera adaptor incorporated during droplet formation and which is adjacent to the 10X cell barcode; in a second PCR, the forward oligo primes on the U6 promoter and adds the Nextera read 2 primer while the reverse oligo primes Nextera read 1 and adds the Illumina P5 adaptor; in the third and last round, a P7 Illumina adaptor and a sample index are incorporated with the forward primer, using the same reverse primer as in the second PCR (see Materials and Methods for more details). These cell-barcoded and sample-indexed sgRNA amplicon libraries can then be sequenced on top of the standard whole transcriptome 10X scRNA-seq libraries to obtain information of the sgRNA expressed in each cell for a high proportion of the cells.

Using the eleven CRISPRa SAM22 cell lines generated above with CROP-sgRNA-MS2 transductions (Figure 3.29, and two non-targeting sgRNA controls), I merged them at equal ratios and ran them through 10X Genomics 3' scRNA-seq droplet formation. After cDNA amplification, I used 10 ng of 10X Genomics barcoded cDNA to perform enrichment PCRs for sgRNA amplicons, following Hills et al. protocol (Hill et al. 2018) (see Materials and Methods). A crucial aspect of this protocol is to monitor PCR amplifications by qPCR to reduce the rate of chimeras that otherwise introduce noise in sgRNA-cell assignments. As a control, I used cDNA from two previous 10X Genomics 3' scRNA-seq libraries of cells transduced with sgRNA(MS2)_puro constructs, whose insertion leads to non- 3'polyadenylated sgRNA transcripts which are not captured during droplet formation with oligo(d)T adaptors. To limit transcript chimera formation, I stopped the PCR reactions at 13 cycles in PCR 1, 12 cycles in

PCR 2 and 10 cycles in PCR 3, before a steady-state of amplification was reached (Figure 3.32). Analysis of these libraries on a bioanalyzer instrument confirmed specific amplification of ~500 bp amplicons (Figure 3.33).

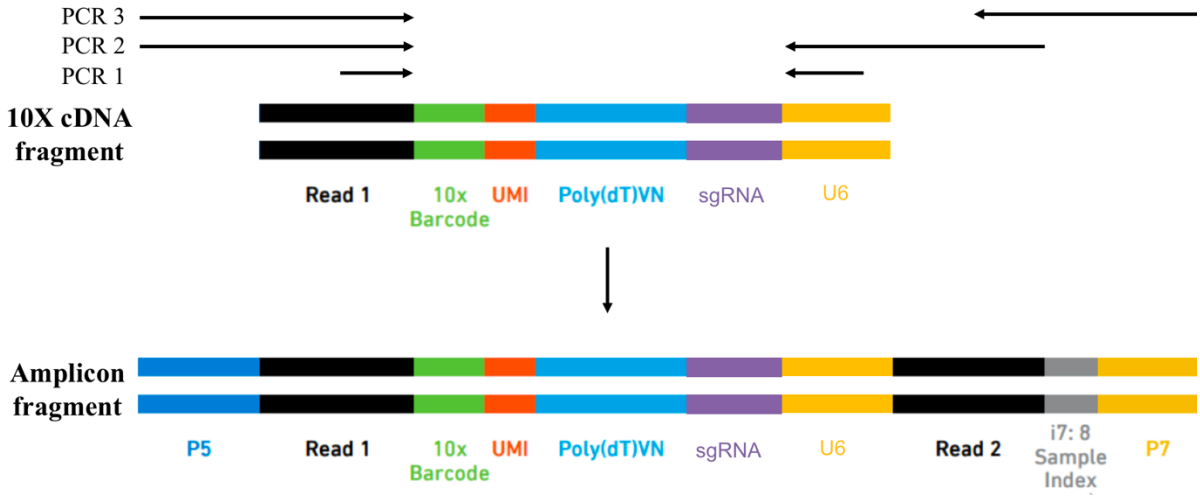



**Figure 3.31– Schematic representation of enrichment PCRs to generate barcoded sgRNA amplicons from 10X Genomics cDNA**

After 10X Genomics 3' scRNA-seq droplet formation, cDNA amplification leads to the formation of fragments (top) that contain a standard Nextera Read 1 adaptor (black), a 10X cell barcode (green), a unique molecular identifier (UMI, red), a poly(d)T sequence (blue), and the 3' end of a transcript, here illustrated as a U6-sgRNA sequence (yellow and purple, respectively) transcribed from the EF1α Pol II promoter in the CROP-sgRNA-MS2 integration. After three rounds of nested PCR amplification, an Illumina P5 adaptor (dark blue) is incorporated at one end of the fragment adjacent to Nextera Read 1 (black, left), and a Nextera Read 2 primer (black, right), an i7 sample index (grey) and an Illumina P7 adaptor (orange) are incorporated at the other end, generating cell-barcoded and sample-indexed sgRNA amplicons ready to be sequenced in an Illumina sequencer (bottom). Picture adapted from 10X Genomics website

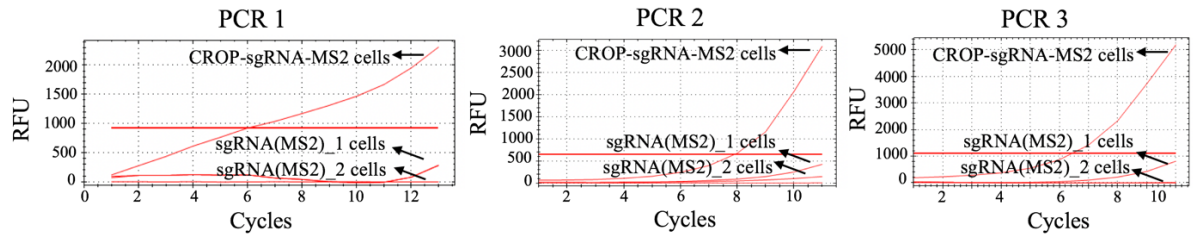



**Figure 3.32– qPCR monitoring of sgRNA enrichment PCRs**

Relative fluorescent units (RFU, y axis) vs number of qPCR cycles (x axis) in each of the three rounds of nested PCRs on barcoded sgRNAs using 10X Genomics cDNA of SAM22 mESCs transduced with CROP-sgRNA-MS2 constructs or with sgRNA(MS2)_puro lentivirus. qPCR was performed with KAPA SYBR and the reactions were stopped before reaching a steady-state of amplification.

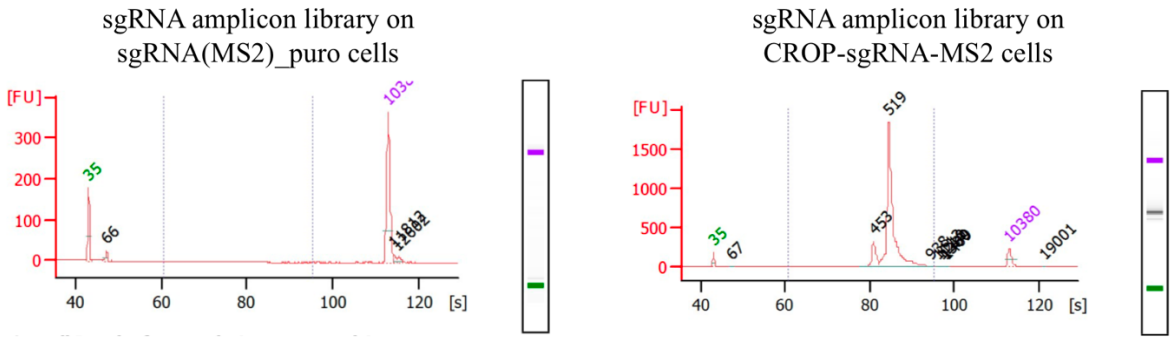



**Figure 3.33– Bioanalyzer traces of sgRNA enrichment libraries**

Bioanalyzer electropherogram showing fluorescence units (FU, y axis) vs migration time in seconds (s, x axis) of sgRNA amplicon libraries of SAM22 mESCs transduced with sgRNA(MS2)_puro constructs (left, control) or with CROP-sgRNA-MS2 lentiviruses (right). Numbers above electropherogram peaks indicate the size in bp of detected fragments, including a lower molecular weight marker of 35 bp (green) and a higher molecular weight marker of 10,380 bp (purple). Only sgRNA amplicon libraries from cells transduced with CROP-sgRNA-MS2 constructs generate amplicons of ~500 bp (right).

Strikingly, after sequencing this CROP-sgRNA-MS2 amplicon sgRNA libraries, 96.45% of the cell barcodes captured could be assigned to expression of a unique sgRNA (Figure 3.34), which is consistent with the percentage reported in Hills et al. 2018 and represents a significant improvement in comparison to the assignment ratio achieved without sgRNA enrichment (Datlinger et al. 2017). With these results, I can conclude that CROP-seq adaptation to the SAM CRISPRa system combined with barcoded sgRNA amplicon libraries is a reliable method to perform a high-throughput pooled CRISPRa screen at scRNA-seq resolution.

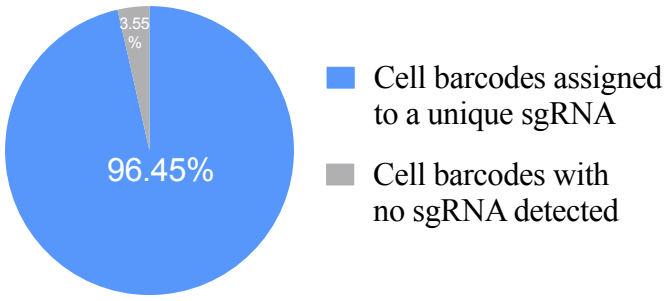



**Figure 3.34– Pie chart of sgRNA-cell assignment from amplicon libraries**

Pie chart showing the percentage of 10X cell barcodes from SAM22 mESCs transduced with CROP-sgRNA-MS2 constructs that could be assigned to expression of a unique sgRNA (blue) as well as the percentage of these cells where a sgRNA could not be detected (grey), after sequencing of sgRNA amplicon libraries.

## 3.3 Conclusions and discussion

In the results presented in this chapter, I first generated mESC lines expressing constitutively the CRISPRa SAM machinery dCas9-VP64 and MS2-p65-HSF1 (Figures 3.2-3.4). Subsequently, I tested the CRISPRa SAM system established in mESCs using MERVL LTRs and *Zscan4* genes as targets, and validated their activation as well as the induction of the expected downstream transcriptional changes by qPCR, western-blot, immunofluorescence and bulk RNA-sequencing (Figures 3.5-3.13). Moreover, I compared the changes induced by *Zscan4* upregulation by either CRISPRa or cDNA overexpression (Figures 3.14-3.16) and also assessed the duration of the transcriptional changes induced by CRISPRa in a time-course experiment (Figures 3.17 and 3.18). Next, I translated the findings from transcriptomic bulk experiments to an scRNA-seq read-out (Figures 3.19-3.26), establishing guidelines and performing calculations critical for subsequent chapters in this dissertation. Lastly, I envisaged a strategy for highly efficient capture of sgRNA sequence information in scRNA-seq experiments where the pool of transduced cells contains information of multiple sgRNAs (Figures 3.27-3.34). Altogether, the experiments described in this chapter set the basis of my dissertation and allowed me to develop an efficient method to do pooled CRISPRa screening at single-cell resolution. In this discussion section, I put these results in context, highlight their relevance in relation to published literature, speculate alternative interpretations and propose further experiments and/or analysis that could help towards a better understanding of the mechanisms regulating CRISPRa in mESCs as well as the MERVL- and ZSCAN4-mediated transcriptional signature in ZGA and 2C-like cells.

Transgenes integrated by lentiviral transfer can be silenced by epigenetic mechanisms (Pfeifer et al. 2002; S. Yao et al. 2004; Herbst et al. 2012). Therefore, it was crucial to generate a mESC line expressing dCas9-VP64 and MS2-p65-HSF1, to be used in subsequent experiments, where the expression of the transgenes was stable. To this end, I generated clonal mESC lines by antibiotic selection and manual colony picking (see Materials and Methods). Whereas dCas9-VP64 integration was more variable across clones, the integration of MS2-p65-HSF1 was efficient in most of the clones tested (Figure 3.3), likely reflecting a higher titer of the lentiviral preparation used for transduction. Moreover, the genomic DNA PCR band for dCas9-VP64

integration in the non-clonal population was very weak (Figure 3.3), consistent with the more variable integration observed across clones and highlighting the importance of generating clonal cell lines. However, in addition to confirming integration and expression of the transgenes, it was crucial to select clones with an unaltered transcriptome for subsequent transcriptomic analysis after transduction with sgRNAs that induce target activation of endogenous loci. RNA-sequencing analysis of two of the clones, SAM7 and SAM22, confirmed high expression of the transgenes and showed that the rest of the transcriptome was largely unchanged (Figure 3.4). These results indicate that the random lentiviral integration of both transgenes and their expression did not interfere with gene expression, and are consistent with previous studies showing that dCas9 by itself does not induce transcriptomic changes (Gilbert et al. 2013; Qi et al. 2013).

MERVL and *Zscan4* are key markers of ZGA and 2C-like cells (Kigami 2002; Peaston et al. 2004; Svoboda et al. 2004; Falco et al. 2007; Zalzman et al. 2010; Macfarlan et al. 2012; Ishiuchi et al. 2015; Ishiguro, Nakatake, et al. 2016; Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018). Therefore, I selected them as targets to optimise my CRISPRa SAM system in SAM22 mESCs, reasoning that if CRISPR-activating them induced a ZGA-like signature, then I could use pooled CRISPRa to screen for multiple regulators. Transduction of two different sgRNAs for each of these targets into SAM22 mESCs led to efficient target gene activation, as analysed by qPCR, immunofluorescence, bulk and single-cell RNA-sequencing (Figures 3.5-3.7, 3.9-3.11, 3.17, 3.21 and 3.22). From these experiments, a key advantage of using CRISPRa over cDNA overexpression became apparent: using a single sgRNA, repetitive elements or multiple genes within a highly similar cluster could be activated at the same time (Figures 3.9-3.11).

As expected, ZGA-like transcription was triggered both by MERVL LTR and *Zscan4* CRISPRa, as seen by qPCR (Figures 3.8 and 3.18), bulk RNA-sequencing (Figures 3.12 and 3.13) and scRNA-seq (Figures 3.24-3.26). Interestingly, higher sgRNA efficiency, measured by activation of the target gene (Figures 3.5, 3.9, 3.10, 3.11 and 3.17), also translated into higher upregulation of the downstream targets (Figures 3.8, 3.12, 3.18), indicating that ZGA-like transcription is dependent on the dose of MERVL and *Zscan4* expression.

ZSCAN4C cDNA overexpression in mESCs was done in the Reik laboratory while I was doing these experiments (Eckersley-Maslin et al. 2019) and very recently also published by another group (W. Zhang et al. 2019). Both studies showed that ZSCAN4C induces ZGA-like expression, same as I observed with CRISPRa of the *Zscan4* gene cluster. However, I showed for the first time MERVL LTR activation (Figures 3.10, 3.21), which triggered transcription downstream of them (Figure 3.11), including the expression of ZGA genes (Figures 3.8, 3.12, 3.13, 3.18, 3.24-3.26). While this was expected, since these repetitive elements can act as functional promoters to drive the expression of hundreds of chimeric ZGA transcripts (Macfarlan et al. 2012; Y. Huang et al. 2017; Franke et al. 2017), genetic manipulation of these elements had not been done before. Furthermore, it revealed a previously unknown inter-regulatory link between *Zscan4* and MERVL, seen consistently in all experiments shown in this chapter (Figures 3.5-3.7, 3.9-3.11, 3.16, 3.17, 3.21 and 3.22). While it is known that both factors belong to the same transcriptional network in the two-cell embryo and they label the same 2C-like subpopulation within mESCs (Kigami 2002; Zalzman et al. 2010; Falco et al. 2007; Macfarlan et al. 2012; Eckersley-Maslin et al. 2016), how they regulate each other has only started to be addressed. ZSCAN4C has been proposed to enhance rather than drive ZGA-like expression in mESCs (Eckersley-Maslin et al. 2019) by direct binding to MERVL LTRs (or MT2 elements), subsequent recruitment of chromatin modifiers and H3K4me1, H3K27ac and H3K14ac deposition on MERVL LTRs that eventually leads to MERVL LTR activation and MERVL expression (W. Zhang et al. 2019). However, my results also show that MERVL LTR activation induces *Zscan4* expression (Figures 3.5-3.7, 3.9-3.11, 3.16, 3.17, 3.21 and 3.22). MERVL LTR elements are present within the *Zscan4* cluster but in opposite orientation to the *Zscan4* genes, making unlikely direct activation of *Zscan4* expression through MERVL LTR elements acting as promoters. It is possible to think of three different mechanisms of how MERVL LTR CRISPRa induces *Zscan4* expression: 1) MERVL LTR CRISPRa recruits chromatin modifications, similar to what direct binding of ZSCAN4C would do (W. Zhang et al. 2019), and subsequently, activated MERVL elements act as enhancers of ZGA transcription, including the *Zscan4* cluster; 2) as seen in my RNA-sequencing results (Figure 3.11), MERVL LTR CRISPRa induces transcription of the genomic regions downstream of these elements, which could result in the upregulation of a direct activator of *Zscan4*; and 3) MERVL LTR CRISPRa provides new transcription factor binding sites (Jichang Wang et al. 2014), which,

indirectly, activates *Zscan4* expression. While beyond the scope of this dissertation, these remain open questions that could be further investigated using this potent tool to activate endogenous transcription of MERVL repeats. For example, MERVL LTR CRISPRa in *Zscan4* KO mESCs (W. Zhang et al. 2019) followed by transcriptome analysis would unravel whether either or both of them are necessary to induce the expression of ZGA genes.

Importantly, I also showed by qPCR that expression of the key pluripotency genes *Oct4* and *Nanog* was not affected after induction of a ZGA-like signature by MERVL LTR or *Zscan4* CRISPRa (Figures 3.8 and 3.18). This is consistent with previous reports showing that protein levels of OCT4 are downregulated in 2C-like cells without alterations at the mRNA level (Macfarlan et al. 2012; Ishiuchi et al. 2015; Eckersley-Maslin et al. 2016; Rodriguez-Terrones et al. 2018).

Despite *Zscan4c* CRISPRa and cDNA overexpression inducing both a ZGA-like transcriptional response, I analysed in more detail the similarities and differences triggered by both methods. PCA clustered CRISPRa and cDNA overexpression samples together and apart from controls, confirming the overall transcriptomic similarities (Figure 3.14). However, differential gene expression analysis showed that, despite a considerable overlap, there were differences in the responses induced by each method (Figure 3.16). Consistent with the qPCR results showing that higher *Zscan4* CRISPRa leads to higher fold changes in the downstream targets (Figures 3.5, 3.8-3.12, 3.17 and 3.18), it is quite likely that the big differences in *Zscan4c* expression levels achieved by CRISPRa and cDNA overexpression (Figure 3.15) explain the transcriptomic differences. In other words, the transcriptional responses might be very similar but, if ZGA-like expression is dependent on the dose of ZSCAN4, there might be some genes that passed the threshold to be called differentially expressed with one method but not the other. In fact, other recent studies comparing cDNA overexpression and CRISPRa showed that downstream responses might be different due to target gene dose effects (Chavez et al. 2015; Sanson et al. 2018; J. Yang et al. 2019). Since CRISPRa triggers transcriptional activation of the endogenous locus, the expression changes of the gene target are possibly more physiologically-relevant than the acute overexpression induced with cDNA transfections. For example, the dosage of OCT4 achieved by either overexpression method influences stem cell reprogramming (J. Yang et al. 2019) and, consistently, CRISPRa has been successfully used

for cellular reprogramming and the study of cellular transitions (Chakraborty et al. 2014; Black et al. 2016; Peng Liu et al. 2018; Weltner et al. 2018; J. Yang et al. 2019). Importantly, there are other technical differences in the application of both methods: 1) for CRISPRa SAM, I harvested cell samples for transcriptomic or protein analysis 6-15 days after sgRNA transduction, whereas ZSCAN4C cDNA overexpression was done over the course of 48 hours; 2) for CRISPRa SAM, transduced cells were selected with antibiotics for expression of the sgRNA, whereas cDNA-eGFP transfected samples from Eckersley-Maslin et al. 2019 were $eGFP^{+/-}$ FACS-sorted before RNA-sequencing analysis. Altogether, the gene dose, the timing and other technical aspects might contribute to the differences seen by *Zscan4* CRISPRa and cDNA overexpression, which, yet, are remarkably similar in the induction of ZGA transcript expression, confirming CRISPRa target specificity. Additionally, CRISPRa presents two extra advantages over cDNA overexpression which are critical in this dissertation: 1) as it does not require cloning of genes, it is highly scalable and allows activation of genes that are, otherwise, difficult to clone or transfect into cells, and 2) by using pooled lentiviral libraries of sgRNAs, many different genes, or other genomic features such as repeat elements (Figures 3.5, 3.10, 3.11, 3.17 and 3.21), can be simultaneously activated within a single pooled experiment, making it a very suitable tool for screening purposes (Konermann et al. 2015; Joung et al. 2017; Horlbeck et al. 2016; Sanson et al. 2018).

To further address the timing effects of the CRISPRa SAM system that I established in mESCs, I performed a time-course experiment, harvesting samples for qPCR analysis 6-15 days after sgRNA transduction into SAM22 cells (Figures 3.17 and 3.18). The earliest time point of 6 days was selected because this was the minimum time needed to transduce sgRNA-expressing lentiviruses into cells and select them effectively with puromycin (see Materials and Methods). Notably, I showed that longer culture of cells, while maintaining puromycin selection, did not affect target gene activation (Figure 3.17) or the expected downstream transcriptional response (Figure 3.18). These results became critical later on with the use of the CROP-sgRNA-MS2 backbone, whose lentiviral titer is reduced compared to the sgRNA(MS2)_puro vector (Datlinger et al. 2017) and, therefore, longer culture times were required for effective lentiviral transduction and antibiotic selection (see Materials and Methods). However, these experiments also raised other questions: how stable are the changes triggered by CRISPRa? does CRISPRa

SAM targeting change the epigenetic state of the targeted promoter? and, if that is the case, are those changes reversible? A study using a doxycyline-inducible CRISPRa system to target the *Nanog* locus in human ESCs showed that *Nanog* activation could be reverted upon doxycycline withdrawal (J. Guo et al. 2017). Nevertheless, larger studies using different cellular systems and different target genes are needed to confirm this observation.

An important observation made in this chapter, using both immunofluorescence and scRNA-seq, was that not all cells selected with antibiotics for sgRNA expression showed target gene upregulation at the RNA (Figures 3.21, 3.22) or protein level (Figures 3.7, 3.31). For example, at the RNA level, approximately 10% of the cells transduced with MERVL LTR sgRNAs activated MERVL expression (Figure 3.21), and 50% of the cells transduced with Zscan4 sgRNAs activated their targets (Figure 3.22). Two alternative hypotheses could explain these observations: 1) antibiotic resistance was acquired by cells that silenced the sgRNA lentiviral construct or were not transduced and, therefore, lack of target gene activation in those cells was due to lack of sgRNA expression; or 2) there was heterogeneity at the single-cell level in the chromatin organisation and/or DNA methylation state of the target gene promoters that allowed the sgRNA to activate transcription in some cells but not in others. The first option seems unlikely given the high percentage of cells that did not activate the target, which is likely too high to be explained by acquisition of antibiotic resistance. However, this possibility could be addressed by analysing expression of the sgRNA at the single-cell level in scRNA-seq libraries. This could not be done in the MERVL and *Zscan4* CRISPRa experiments presented in this chapter since the scRNA-seq libraries were performed on cells transduced with the sgRNA(MS2)_puro construct, which produces a sgRNA transcript that is not captured by the oligo(d)T-conjugated beads in 10X Genomics 3' scRNA-seq libraries. It will be addressed in the following chapter with cells transduced with CROP-sgRNA-MS2 lentiviruses. The second hypothesis is also tested in chapter 4, using scNMT-seq (single-cell NOMe, methylation and transcription-sequencing) (Clark et al. 2018), a technique that allows profiling chromatin accessibility, DNA methylation and transcription from the same single cell.

On a more technical aspect, the cell capture efficiency obtained in the scRNA-seq libraries presented in this chapter was considerably low (approximately 25-30%, depending on the sample), probably due to the quality of the sample preparation, reflected in a population of

"low-quality" cells (Figure 3.20). The sample preparation procedure was improved upon in subsequent experiments (see Materials and Methods).

More importantly, this scRNA-seq experiment allowed a power calculation to be made using the percentage of cells that activated a ZGA-like signature upon MERVL and *Zscan4* activation. While these calculations are critical to design the multiplexed pooled screen (see chapter 4), it is essential to keep in mind that not all regulators of ZGA might act via the MERVL/*Zscan4*-regulated transcriptional pathway and, therefore, more or less cells than the 367 calculated per sgRNA might be needed to detect a screen hit with enough statistical power.

In the last result section of this chapter, I decided to use CROP-seq (Datlinger et al. 2017) as a strategy to capture sgRNA sequence information in scRNA-seq libraries over the published alternative Perturb-seq or CRISP-seq strategies (Jaitin et al. 2016; Dixit et al. 2016; Adamson et al. 2016). Whereas CROP-seq uses a clever cloning strategy to read directly the sgRNA sequence from a polyadenylated transcript by exploiting the replication cycle of lentivirus into the host genome (Datlinger et al. 2017) (Figure 1.18), Perturb-seq- and CRISP-seq-based approaches use expression vectors to co-express polyadenylated barcodes alongside the non-polyadenylated sgRNAs; these barcodes can be detected by scRNA-seq and used to assign sgRNA expression to each individual cell (Jaitin et al. 2016; Dixit et al. 2016; Adamson et al. 2016). However, since Pertub-seq/CRISP-seq was published, it has been shown that this strategy often leads to inaccurate sgRNA-cell assignment in up to 50% of the cells due to uncoupling of the sgRNA-barcode pairs during lentiviral packaging of pooled libraries (Hill et al. 2018; S. Xie et al. 2018; Adamson et al. 2018; Feldman et al. 2018). This could be overcome by arrayed packaging of lentiviral libraries, which drastically reduces the throughput and scalability of screening experiments. Even though some studies have proposed alternatives to use Perturb-seq in lentiviral pooled libraries (Adamson et al. 2018; Feldman et al. 2018), the CROP-seq strategy completely abolishes the problem of barcode uncoupling, making it more suitable for pooled CRISPR screening (Hill et al. 2018). However, CROP-seq does not come without constraints. As mentioned earlier, disruption of the 3' LTR with the U6-sgRNA cassette reduced the titer of the lentivirus (Datlinger et al. 2017) (see Materials and Methods), and, consequently, cloning more than one sgRNA in the CROP-seq vector to introduce multiple perturbations in a cell is not feasible, unless using high MOI transductions to introduce multiple

random combinatorial perturbations in a cell. Recently, an alternative strategy was proposed to capture directly the non-polyadenylated sgRNA sequence in 10X Genomics scRNA-seq libraries using a specialised barcode introduced in the beads together with oligo(d)T adaptors (Replogle et al. 2018). This strategy combines the best features of each method and is likely to be implemented widely for CRISPR screening at single-cell resolution.

I adapted the CROP-seq vector (Datlinger et al. 2017) to include MS2 loops in the scaffold sequence of the sgRNA and allow recruitment of MS2-p65-HSF1 protein in SAM22 cells (Figure 3.27). The sgRNA expressed from this CROP-sgRNA-MS2 vector could be captured by oligo(d)T adaptors (Figure 3.28), simulating capture of polyadenylated transcripts in 10X Genomics 3' scRNA-seq droplets, and did not impair CRISPRa of endogenous targets (Figures 3.29 and 3.30). Moreover, specific amplification of sgRNA-containing fragments in scRNA-seq libraries (Hill et al. 2018) showed that approximately 96% of the cell barcodes recovered could be assigned to a unique sgRNA (Figure 3.31-3.34). Therefore, CROP-seq adaptation to CRISPRa SAM coupled to sgRNA amplification promises to be a highly efficient method to use and assign sgRNA expression to each individual cell in the pooled screen for regulators of ZGA (see chapter 4). However, it is important to consider that the 96% unique assignment rate might be optimistic for two reasons: 1) it was calculated based on captured cell barcodes in the amplicon libraries which were not compared to the total number and quality of cells captured in the whole-transcriptome libraries; and 2) I did individual sgRNA transductions and only pooled cells at the time of sequencing, therefore discarding the possibility of infecting a cell with multiple different sgRNAs.

Lastly, in addition to CRISPRa SAM, I anticipate the CROP-sgRNA-MS2 vector could also be used in multiple other CRISPR editing strategies with scRNA-seq read-out, by fusing the MS2 RNA binding protein to other epigenetic modifiers (see section 1.4.1).

# Chapter 4 Single-cell transcriptomics CRISPR activation screen for regulators of ZGA

## 4.1 Background and summary

To identify regulators of ZGA, one could perform a flow-cytometry-based screen using ZGA reporters, such as MERVL or *Zscan4* (Macfarlan et al. 2012; Zalzman et al. 2010; Eckersley-Maslin et al. 2016), as previously done in recent studies (Rodriguez-Terrones et al. 2018; Eckersley-Maslin et al. 2019; X. Fu et al. 2019; Y.-L. Yan et al. 2019). However, transcriptomic analysis have shown that ZGA is a complex transcriptional process and, additionally, the MERVL/*Zscan4*-regulated network only comprises a small fraction of all the genes that are transcribed during ZGA (Xue et al. 2013; Deng et al. 2014; S.-J. Park et al. 2015; Eckersley-Maslin et al. 2016). Therefore, relying on a few markers for screening might hamper the identification of crucial ZGA regulators. Pooled CRISPRa screening at single-cell transcriptomic resolution represents an ideal system for high-throughput identification of regulators of complex transcriptional activation process, such as ZGA, by providing transcriptome-wide insights into the changes triggered by potential hits. However, while powerful in its read-out, scRNA-seq methods are limited by the sequencing costs per cell and, therefore, a targeted approach is required in the selection of the screen candidates.

ZGA occurs as part of the MZT, a developmental transition in which there is not only a transcriptional handover from the mother to the embryo, but also extensive remodeling of the epigenetic and chromatin landscapes (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018; Jansz & Torres-Padilla 2019; Ladstatter & Tachibana 2019) (see section 1.1). With these premises, I hypothesised that maternal epigenetic and transcriptional factors play a crucial role in ZGA regulation.

In this chapter, I first defined a list of screen candidates consisting of maternal proteins with epigenetic and transcriptional roles. A pooled sgRNA lentiviral library targeting the promoters of these candidates was designed and constructed and, after titration, transduced into SAM22 mESCs. scRNA-seq of more than 300,000 cells, together with amplicon sgRNA libraries, allowed analysis of the sgRNA expressed in each cell and, consequently, analysis of target gene activation. This large dataset provided an opportunity to systematically identify features that contribute to CRISPRa efficiency. I evaluated both sgRNA and target gene features to understand, on the one hand, what makes an efficient sgRNA and, on the other hand, whether some genes have transcriptional or epigenetic patterns that make them more prone to upregulation by CRISPRa. I also investigated CRISPRa efficiency at the single-cell level since both the results from this screen and previous results presented in chapter 3 revealed that even with all the CRISPRa machinery active and functional sgRNAs, some cells are resistant to activate the target gene.

## 4.2 Results

### 4.2.1 Selection of candidates and screen design

Given the extensive epigenetic and transcriptional changes that occur during the MZT (reviewed in Eckersley-Maslin, Alda-Catalinas & Reik 2018; Jansz & Torres-Padilla 2019; Ladstatter & Tachibana 2019), I decided to screen for maternal factors with epigenetic and transcriptional regulation roles. For that, I looked into three proteomic studies of mouse oocytes and zygotes (Pfeiffer et al. 2011; Pfeiffer et al. 2015; B. Wang et al. 2016). The Pfeiffer et al. 2011 dataset analysed the proteome of MII mouse oocytes to a depth of 3,699 proteins using label-free liquid chromatography tandem-mass spectrometry (LC-MS/MS); the Pfeiffer et al. 2015 dataset used stable isotope labelling with amino acids in cell culture (SILAC) in conjunction with LC-MS/MS to detect 2,010 proteins present simultaneously in MII oocytes of four inbred strains (129/Sv, C57Bl/6J, C3H/HeN, DBA/2J); and the B. Wang et al. 2016 dataset detected 5,350 proteins in MII oocytes and 5,448 in zygotes using SILAC LC-MS/MS, for which I selected 2,897 detected both in MII oocytes and zygotes. Next, I used gene ontology (GO) to identify 1,683 genes with roles in transcription, chromatin and epigenetic regulation

in any given biological context (see Materials and Methods) and, by intersecting this GO list with each proteomic dataset, I obtained a list of 230 screen candidates (Figure 4.1 and Appendix C). These included 22 positive controls of proteins previously described to play a role in ZGA regulation, either *in vivo* – HSF1 (Christians et al. 2000), SMARCA4 (Bultman et al. 2006), CTCF (Wan et al. 2008), SOX2 (Pan & Schultz 2011), NFYA (F. Lu et al. 2016), LIN28A (K.-H. Kim et al. 2016), YAP1 (C. Yu, Ji, Dang, et al. 2016), DUX (Hendrickson et al. 2017; De Iaco et al. 2017; Iaco et al. 2019; Z. Chen & Yi Zhang 2019) and KDM5B (Dahl et al. 2016; B. Zhang et al. 2016) – or shown to upregulate a ZGA-like signature in mESCs – MERVL, ZSCAN4 (Macfarlan et al. 2012; Zalzman et al. 2010; Eckersley-Maslin et al. 2016; Eckersley-Maslin et al. 2019, W. Zhang et al. 2019, and results presented in chapter 3), BAHD1, EYA1, IRF1, GATA3, TOX3, TRP63, HDAC9, USP3, TET3, DPPA2 and DPPA4 (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Y.-L. Yan et al. 2019) (Figure 1.12). Moreover, six factors tested in a previous study in mESCs that did not upregulate a ZGA-like network, namely PLAC8, LHX2, CALCOCO1, RB1, SS18L1 and NPM2 (Eckersley-Maslin et al. 2019), were included as negative controls together with four genes not expressed in oocytes or during pre-implantation development (LYG2, SPATA3, GPR6 and FAM17A).

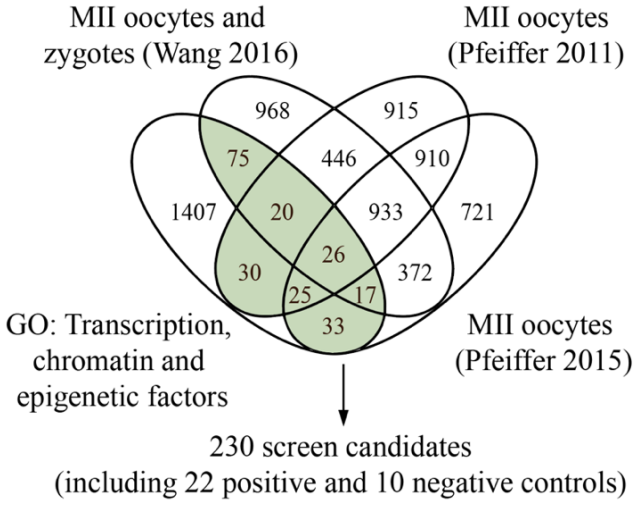



**Figure 4.1– Selection of screen candidates: maternal transcription factors and epigenetic regulators**

The 230 screen candidates were selected from three proteomic datasets of metaphase II (MII) mouse oocytes and zygotes (Pfeiffer et al. 2011; Pfeiffer et al. 2015; B. Wang et al. 2016) using gene ontology (GO) enrichment of transcription, chromatin, epigenetic regulation. Factors detected in any of the three datasets with a relevant GO term were selected, together with 22 positive and 10 negative controls.

Not only at the proteome level, but also at the mRNA level, these candidates are expressed in oocytes, zygotes and during early stages of pre-implantation development (Figure 4.2). Importantly, their expression is not affected by constitutive expression of dCas9-VP64 and MS2-p65-HSF1 in mESCs (Figure 4.3), consistent with the largely unaltered transcriptome of SAM22 mESCs compared to E14 (Figure 3.4).

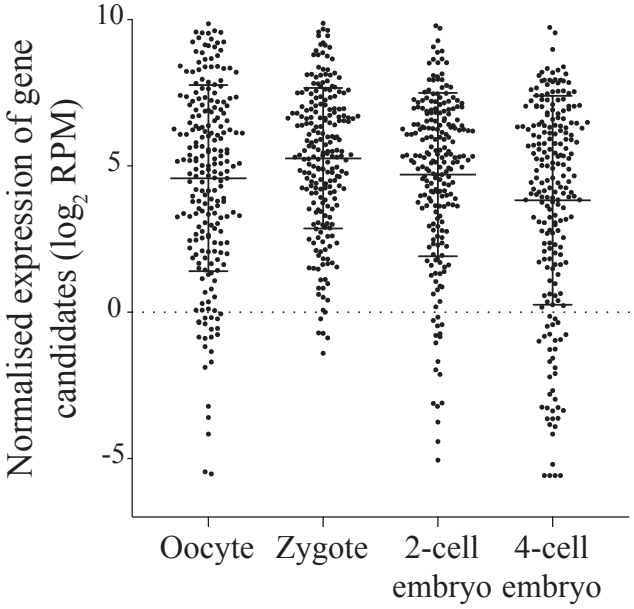



**Figure 4.2– Expression of screen candidates in oocytes, zygotes, 2-cell and 4-cell embryos**

Dot-plots showing normalised expression in $\log_2$ reads per million (RPM) of screen candidates in mouse oocytes, zygotes, two-cell and four-cell embryos (data analysed from Xue et al. 2013).

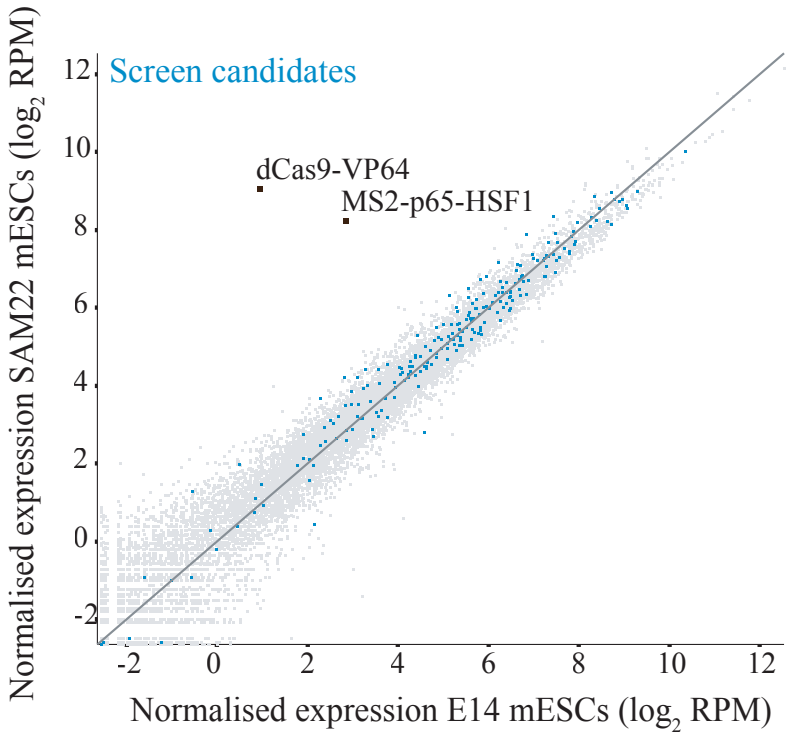



**Figure 4.3– Expression of screen candidates in SAM22 and E14 mESCs**

Scatterplot showing normalised gene expression in $log_2$ reads per million (RPM) of SAM22 mESCs against E14 mESCs, highlighting screen candidates in blue and dCas9-VP64 and MS2-p65-HSF1 transcripts in black.

Despite the considerable reduction of the costs of scRNA-seq, especially with droplet-based automated technologies such as 10X Genomics, the sequencing cost per cell still represents an important limitation and, therefore, a carefully considered experimental design is needed. First, to account for ineffective sgRNAs, at least two sgRNAs per target gene should be included in the pooled library, alongside multiple non-targeting sgRNA controls (Figure 4.4). In a preliminary experiment (section 3.2.3), I determined that, in order to have the statistical power to detect a ZGA-like signature in a positive screen hit, a minimum of 367 cells had to be sequenced per sgRNA. Furthermore, the efficiency for assigning sgRNA expression to a cell is approximately 96% (Figure 3.34), that is, without accounting for multiple infection events that can occur with a pooled lentiviral library, 4% of the cells sequenced would have to be discarded. Altogether, these calculations suggested that at least ~181,600 cells would need to be sequenced to cover the 230 perturbations, each with 2 sgRNAs, with enough statistical power (Figure 4.4). A 10X Genomic chip contains 8 lanes, each of which can capture up to

10,000 single cells. Consequently, after delivery and selection of the pooled sgRNA library, cells would need to be sequenced across at least 19 10X Gennomics lanes, that is, two full chips and three extra lanes from a third chip. To increase statistical power, and considering that not all ZGA regulators might act via the MERVL/*Zscan4*-regulated transcription network used to perform power calculations, I decided to sequence cells across three full 10X Genomics chips (Figure 4.4).

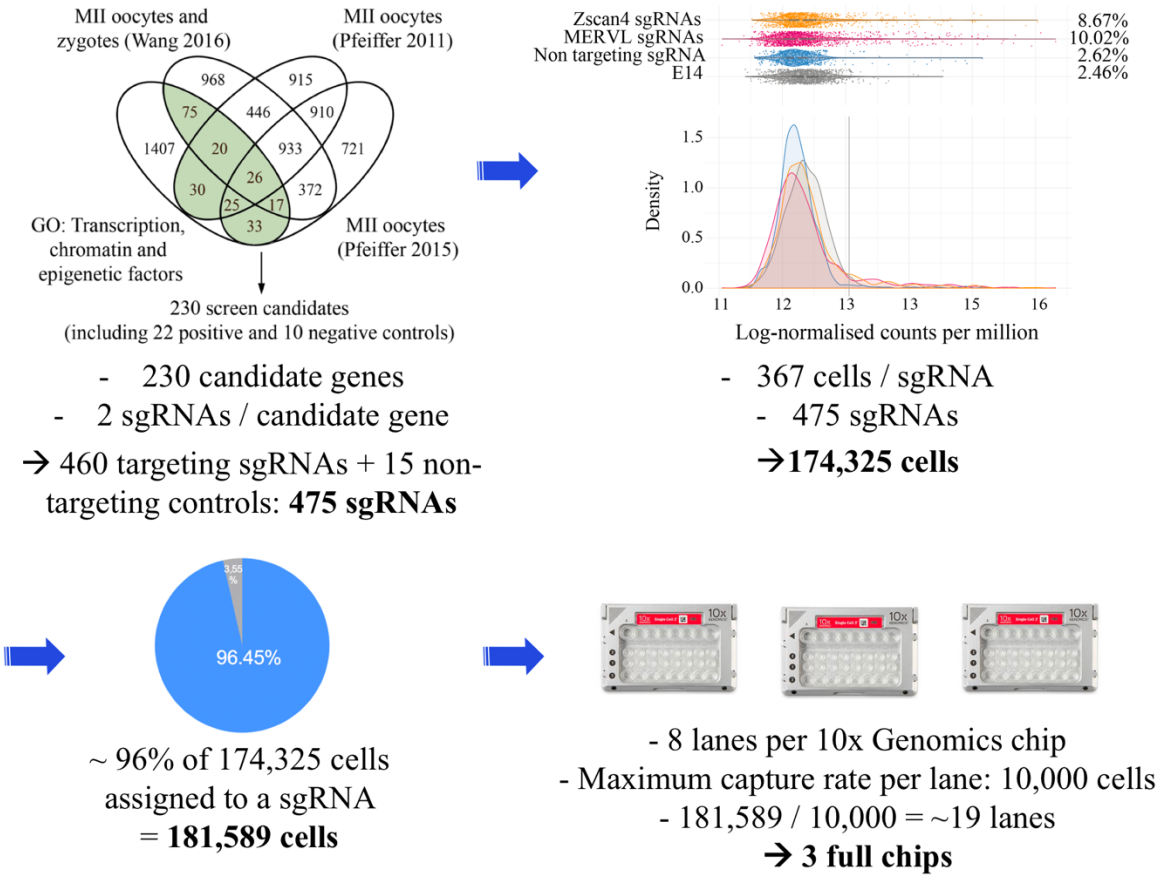



**Figure 4.4– Work-flow to determine the number of cells to be sequenced**

The 230 screen candidates will be targeted with two sgRNAs each. Additionally, 15 non-targeting sgRNA controls will be included, constituting a pooled lentiviral library of 475 sgRNAs. To cover each sgRNA with 367 cells, based on previous power calculations, 174,325 cells would be needed. Additionally, the efficiency of sgRNA detection in a cell is approximately 96%, which increases the number of cells to be sequenced to 181,589. Up to 10,000 cells can be captured per 10X Genomics chip lane, that is, to sequence 181,589 cells, 19 lanes would be needed (more than two chips). To increase statistical power, a final decision was made to sequence cells across 3 full chips.

### 4.2.2 Design, construction and transduction of a custom sgRNA lentiviral library

Using the genome-wide CRISPRa library designed by Joung and colleagues (Joung et al. 2017), I designed a custom oligo library containing two sgRNAs per candidate gene with flanking sequences for cloning into the CROP-sgRNA-MS2 construct, which was then assembled by VectorBuilder (see Materials and Methods). I selected the top two sgRNAs for each target gene ranked based on on-target and off-target scores (Joung et al. 2017). These sgRNAs were targeted to the 180 bp window upstream of the TSS of each candidate gene. Importantly, for genes with multiple isoforms, I selected sgRNAs that would target the oocyte-specific isoform, based on expression data from Veselovska et al. 2015. Fifteen non-targeting sgRNA controls were also included, as design in Joung et al. 2017, constituting a library of 475 sgRNAs. After cloning into the CROP-sgRNA-MS2 vector, next-generation sequencing confirmed representation of all sgRNAs in the cloned oligo library (Figure 4.5, Appendix C).

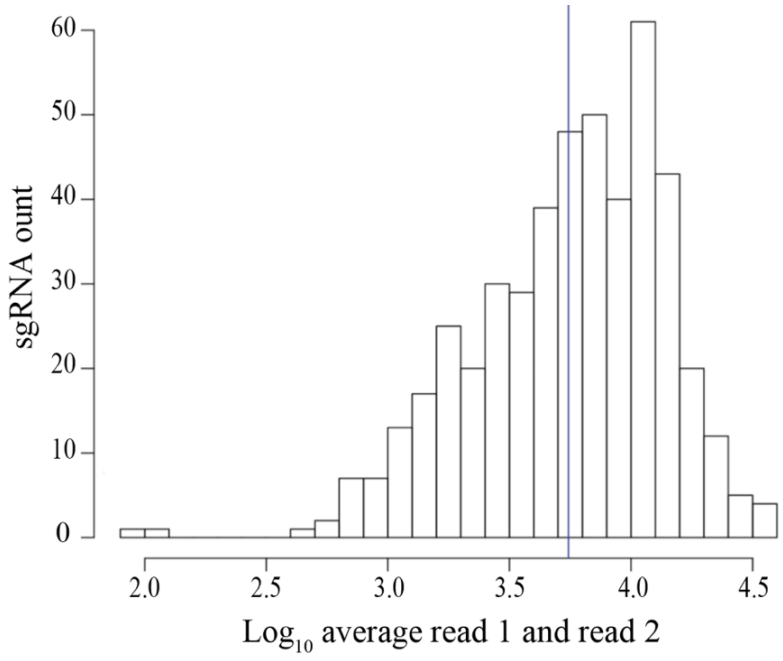



**Figure 4.5– Histogram showing the distribution of sgRNA representation in the cloned oligo library**

Histogram of sgRNAs showing normalised $\log_{10}$ average of read 1 and read 2 from next-generation sequencing analysis of the 475 oligo sgRNA library cloned into the CROP-sgRNA-MS2 lentiviral construct. The blue line represents a mean of 3.74 $\log_{10}$ average of read 1 and read 2; the standard deviation from the mean is 0.39; representation of 90% of sgRNAs in the library is within 19.9 folds.

After lentiviral packaging, transduction of this pooled sgRNA library at low MOI (number of lentiviral particles infecting each cell) was crucial to ensure the introduction of a unique perturbation in each cell and avoid the confounding effects of multiple perturbations or activation of multiple genes. The number of integration events achieved by lentiviral transduction follows a Poisson distribution (Fehse et al. 2004), which suggests that transduction at a MOI of 0.1 reduces the probability for every cell to acquire two or more infection events to 0.47% (Figure 4.6). Therefore, I transduced different volumes of the lentiviral sgRNA library into 120,000 SAM22 mESCs and analysed mCherry expression by flow cytometry analysis (Figure 4.7). As mCherry was included in the CROP-sgRNA-MS2 vector (Figure 3.27), this experiment allowed titration of the library and calculation of a simple equation by linear regression to estimate the percentage of transduced cells with different viral volumes (Figure 4.8).

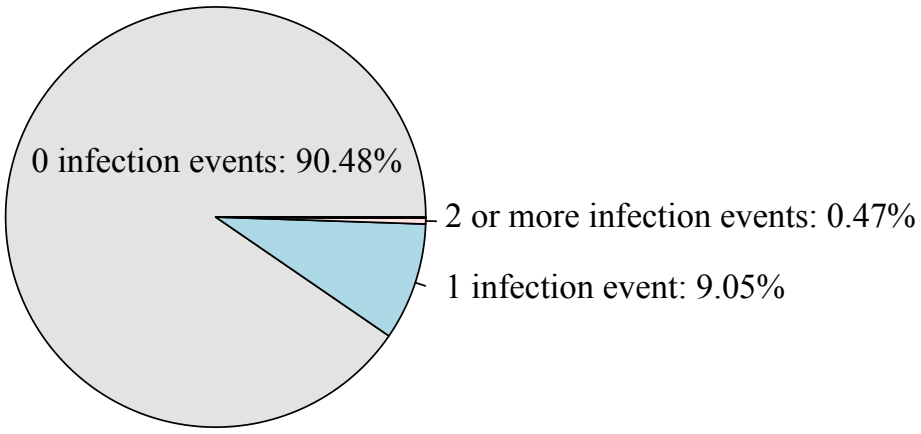



**Figure 4.6– Pie chart showing probability of number of infection events at a MOI of 0.1**

Pie chart showing the probability of a cell being transduced with 0, 1, 2 or more lentiviruses when the transduction is performed at a MOI of 0.1, as calculated by Poisson distribution.

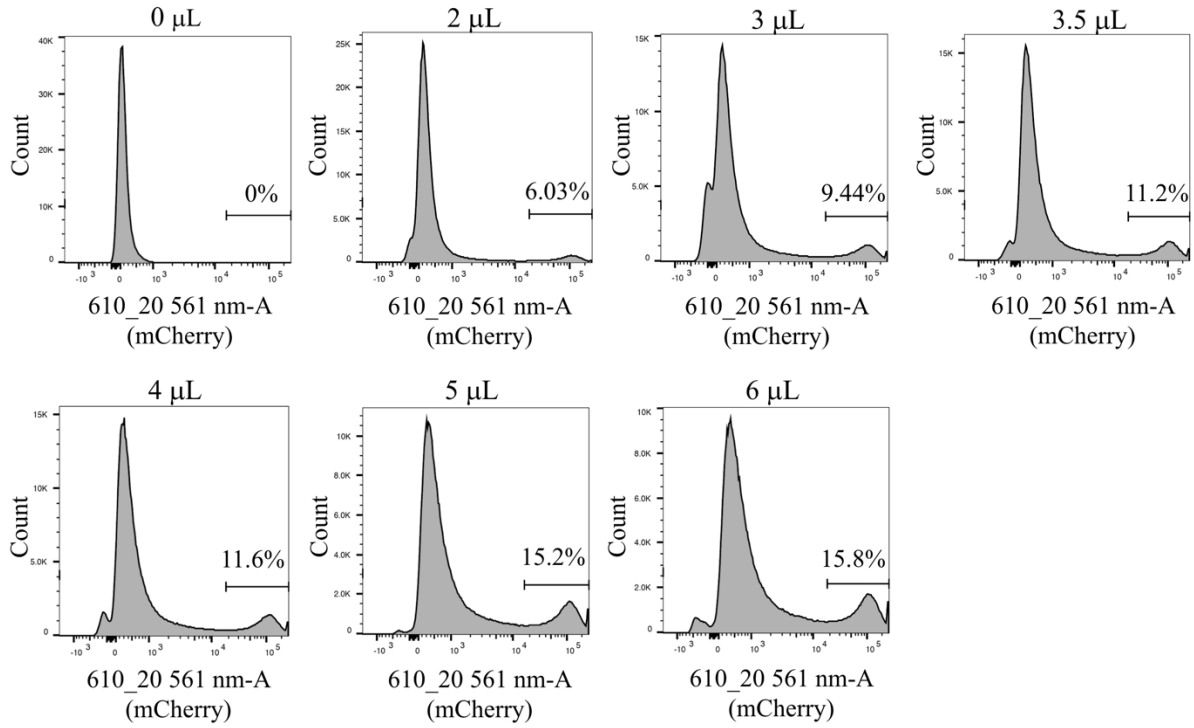



**Figure 4.7– mCherry histograms for library titration analysed by flow cytometry**

Histograms showing expression of mCherry in SAM22 mESCs transduced with different volumes of the 475 sgRNA lentiviral library, analysed by flow cytometry two days after transduction; percentages of cells transduced (expressing mCherry) are shown.

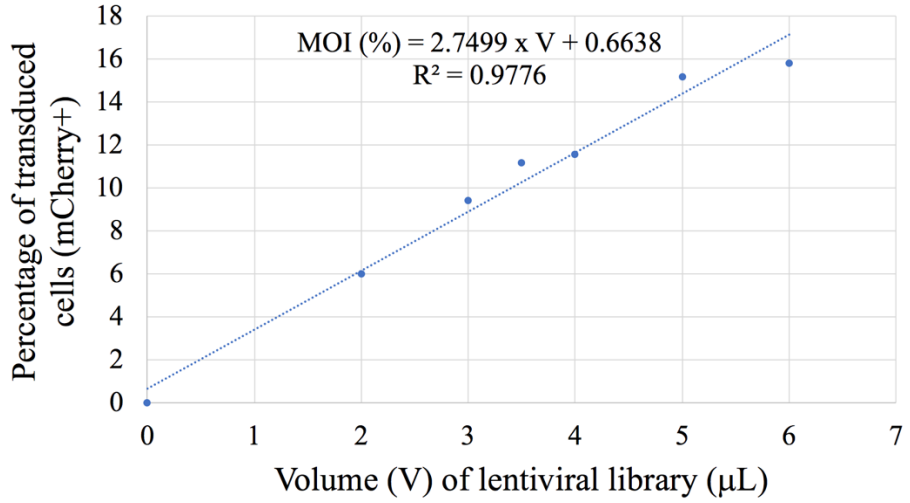



**Figure 4.8– Calculation of lentiviral MOI equation**

Scatterplot showing the percentage of transduced SAM22 mESCs (mCherry$^+$) with different volumes (V) of the 475 sgRNA lentiviral library. From the data, a linear regression equation was calculated to allow estimation of the multiplicity-of-infection (MOI) as percentage of transduced cells, as a function of lentiviral volume in µL.

With this formula, I could estimate that approximately 3.4 µL were required to transduce 120,000 SAM22 mESCs at a MOI of 0.1. In other words, transduction with 3.4 µl of lentiviral library into 120,000 cells would lead to 10% of transduced cells, most of them likely with a unique lentivirus. I aimed to have a coverage of at least 1,000 cells for each of the 475 sgRNAs in the library at the time of transduction to minimise potential proliferation biases and guarantee sufficient representation of each perturbation at the time of scRNA-seq. Accounting for that and for a MOI of 0.1 (10% of transduced cells), I transduced 4.75 million SAM22 mESCs with 135 µl of lentiviral library, in triplicate.

$$\text{For 120,000 cells: } MOI\ (\%) = 2.75\ x\ V + 0.6638\ ;$$

$$10 = 2.75\ x\ V + 0.6638 \rightarrow V = 3.4\ \mu l$$

$$475\ sgRNAs\ x\ 1000\ \frac{cells}{sgRNA} = 475{,}000\ transduced\ cells$$

$$475{,}000\ x\ 10 = 4.75\ million\ cells\ to\ plate\ for\ transduction$$

$$\frac{3.4\ \mu l}{120{,}000\ cells} = \frac{x\ \mu l}{4{,}750{,}000\ cells} \rightarrow x = 134{,}583\ \mu l$$

Flow cytometry analysis two days after transduction, before starting puromycin selection, revealed that less than 10% of the cells were transduced (Figure 4.9), which ensured that after antibiotic selection, most remaining cells would have a unique single perturbation.

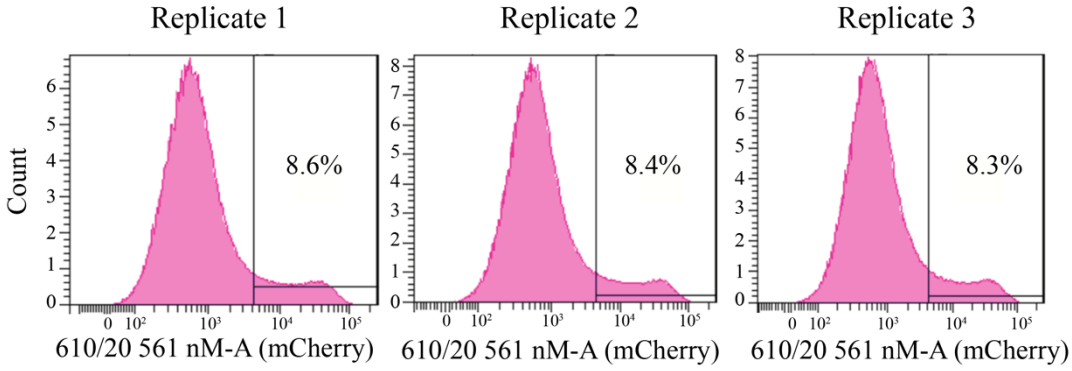



**Figure 4.9– mCherry histograms of three screen replicates analysed by flow cytometry**

Histograms showing expression of mCherry in three replicates of SAM22 mESCs transduced with the 475 sgRNA lentiviral library at low MOI, analysed by flow cytometry two days after transduction; percentages of cells transduced (expressing mCherry) are shown.

After selecting and expanding the CRISPRa-perturbed cells for 10 days (Figure 4.10) (see Materials and Methods), 160,000 cells from each replicate were loaded across 8 lanes of a 10X Genomics chip, aiming to recover 80,000 single-cell transcriptomes for each replicate. While having previously calculated that a minimum of 367 cells per sgRNA were required to have enough statistical power, I decided to perform the lentiviral transduction three times to account for technical variability and sequence less cells per sgRNA in each individual replicate, reasoning that, if there were no major batch effects between replicates, they could be pooled to reach statistical power.

After running the cells through the 10X Genomics chromium controller, 10 ng of barcoded cDNA were used to perform sgRNA amplicon libraries, as described in chapter 3, and 700-1100 ng were used for whole-transcriptome library preparation (Figure 4.10) (see Materials and Methods).

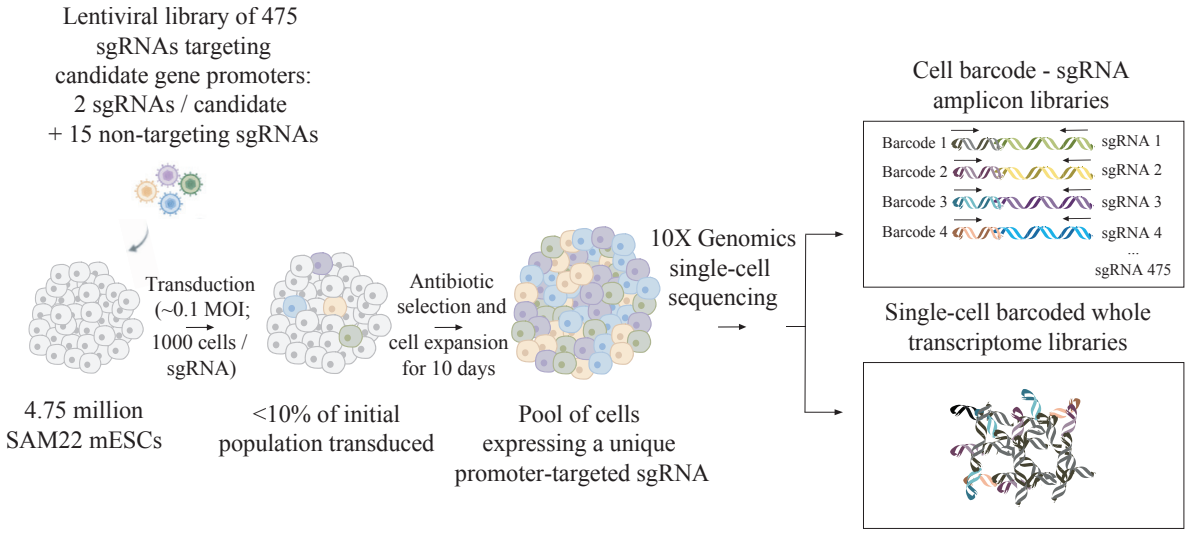



**Figure 4.10– Schematic overview of the CRISPRa scRNA-seq screen**

Schematic overview of the screen showing transduction of a lentiviral library of 475 sgRNAs at ~0.1 MOI into SAM22 mESCs, aiming for a representation of a 1000 cells per sgRNA, considering only ~10% of the cells would be transduced; the cells were selected with puromycin, encoded in the CROP-sgRNA-MS2 vector, and expanded for 10 days in culture before being sequenced using 10X Genomics 3' single-cell RNA-sequencing technology; full transcriptome libraries as well as barcoded sgRNA amplicon libraries were generated.

### 4.2.3 Quality control of single-cell RNA-sequencing and sgRNA assignment

Each 10X whole transcriptome scRNA-seq library corresponding to a 10X Genomics chip lane was sequenced on an Illumina HiSeq4000 lane and the amplicon sgRNA libraries were multiplexed and run across two HiSeq2500 lanes (see Materials and Methods). A total of 341,103 cells were captured (Figure 4.11), which represents a capture efficiency of 71%.

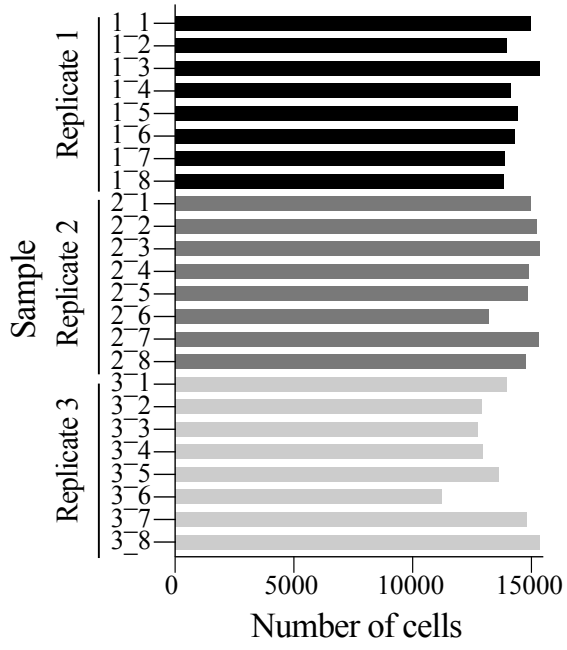



**Figure 4.11– Number of cells captured in each 10X Genomics scRNA-seq library**

Number of cells captured in each 10X Genomics scRNA-seq library out of the 20,000 loaded SAM22 mESCs transduced with the 475 sgRNA lentiviral library. Each transduction replicate (1, 2 or 3) was run across 8 lanes (1 to 8).

For each of the transduction replicates, cells with less than 4,000 or more than 20,000 unique molecular identifiers (UMIs), cells with less than 1,600 or more than 5,000 detected genes and cells with more than 5% reads coming from mitochondrial genes were filtered-out to discard low-quality input and cell doublets (Figure 4.12) (see Materials and Methods). A total of 317,847 cells across the three replicates (93,2%) passed this quality control. Next, sgRNA expression was assigned to each single cell using the barcoded sgRNA amplicon libraries (see Materials and Methods). Out of the 317,847 cells that passed quality control, 249,676 cell barcodes (78.6%) were captured in the amplicon libraries. Out of these, a total of 203,894 cells

across the three replicates (81.66%) were assigned to a unique sgRNA, while the remaining cells showed expression of two (3.73 %) or more sgRNAs (14.15%) or could not be confidently assigned to a sgRNA in the pooled library (0.46%) (Figure 4.13) (see Materials and Methods). These percentages and their significance are discussed in detail in section 4.3 of this chapter. For downstream analysis, only cells with expression of a single sgRNA were taken forward to avoid potential confounding effects of multiple sgRNAs.

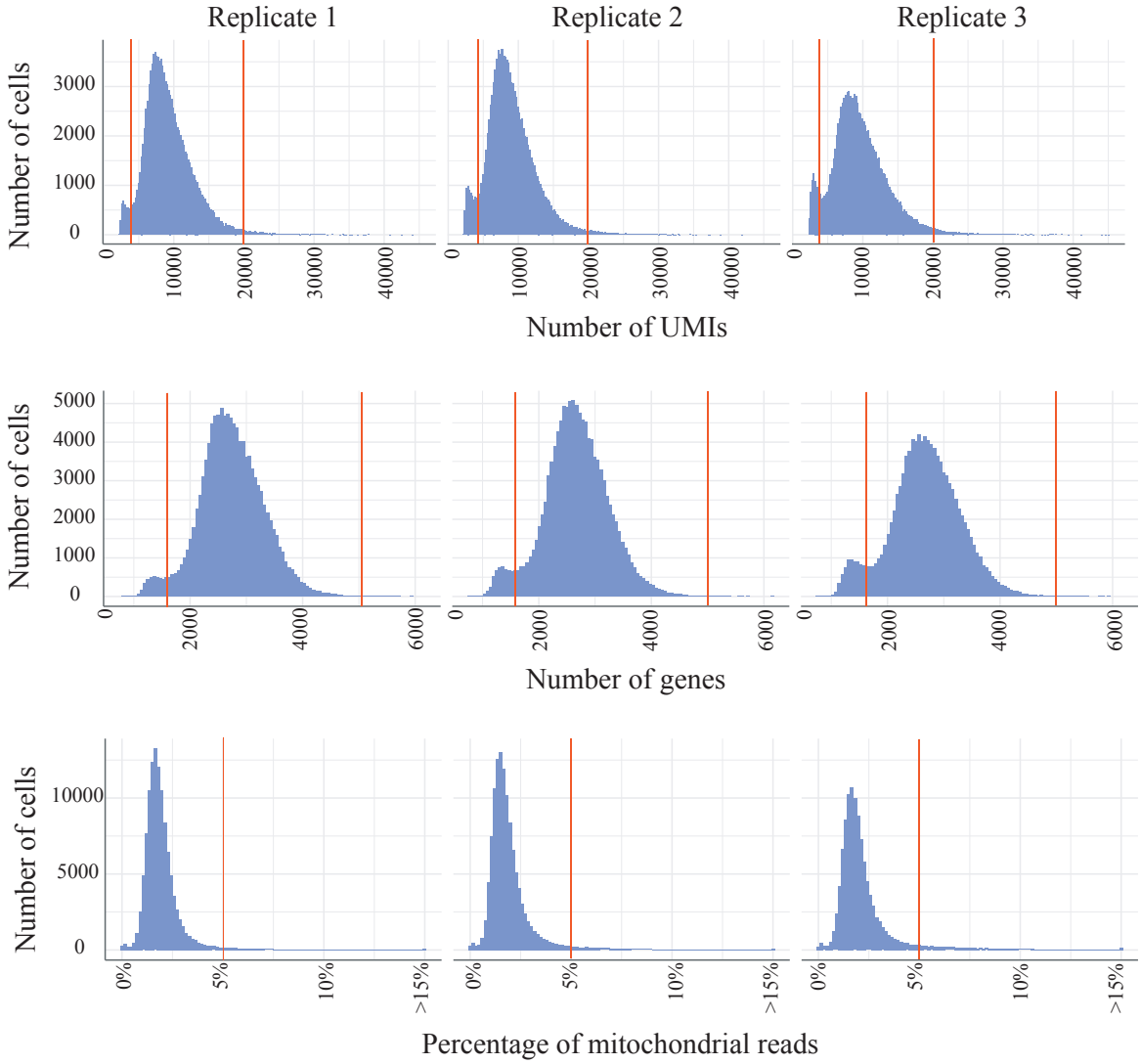



**Figure 4.12– Quality control and filtering of 10X Genomics scRNA-seq libraries**

Histograms showing number of unique molecular identifiers (UMIs) (top panel), number of genes detected (middle panel) and percentage of reads from mitochondrial genes (bottom panel) in each of the transduction replicates (1-3). As indicated with red vertical lines, cells with more than 4,000 and less than 20,000 UMIs (top panel), more than 1,600 and less than 5,000 genes (middle panel) and less than 5% of mitochondrial reads (bottom panel) were retained as high-quality cells.

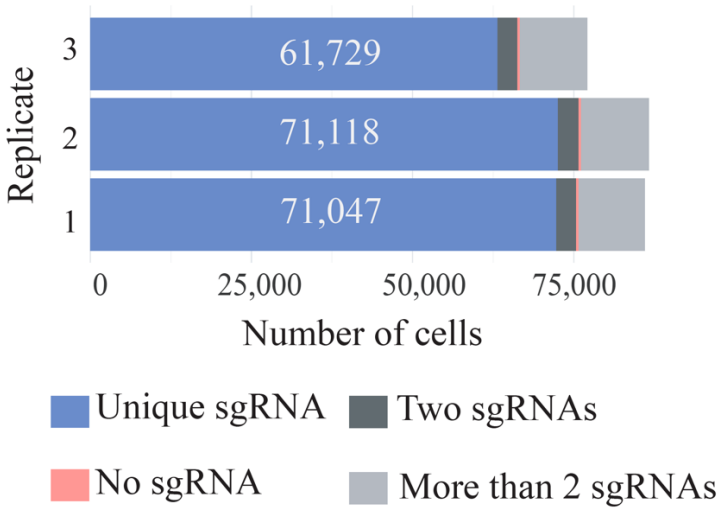



**Figure 4.13– Number of cells assigned to sgRNA expression**

Number of cells assigned to a single unique sgRNA (blue), two sgRNAs (dark grey), more than two sgRNAs (light grey) or none (pink) in each of the three transduction replicates (1-3). The number of cells assigned to a unique sgRNA in each replicate are displayed in white.

Given that not all the sgRNAs were equally represented in the sgRNA plasmid library (Figure 4.5), I next analysed the number of cells transduced and captured with each sgRNA (Appendix C). The distribution of number of cells per sgRNA was almost identical between replicates (Figure 4.14) and, after merging them, the dataset consisted of 437 cells per sgRNA on average (Figure 4.15), which was above the power estimate of a minimum of 367 cells per sgRNA to detect a ZGA signature. All sgRNAs cloned into the lentiviral library were represented in the 10X Genomics scRNA-seq dataset (Appendix C). The number of cells captured expressing each sgRNA ranged from 9 to 3,300 (Appendix C) and correlated with sgRNA representation in the plasmid library before transduction, indicating that activation of the target genes did not have any strong effects on cell proliferation or viability (Figure 4.16).

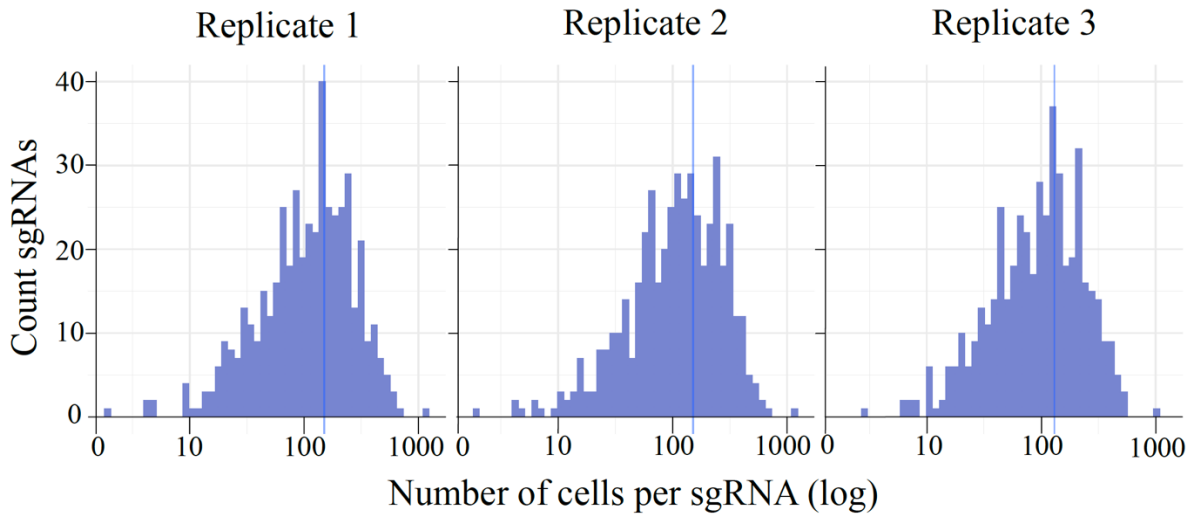



**Figure 4.14– Histograms of sgRNA count in relation to number of cells per sgRNA, in each replicate**

Histograms of the number of cells expressing a given unique sgRNA from the 475 sgRNA library in each transduction replicate (1-3); the number of cells per sgRNA is depicted in logarithmic scale and the vertical blue line represents the average of cells per sgRNA (150 cells in replicate 1, 151 cells in replicate 2, and 131 cells in replicate 3).

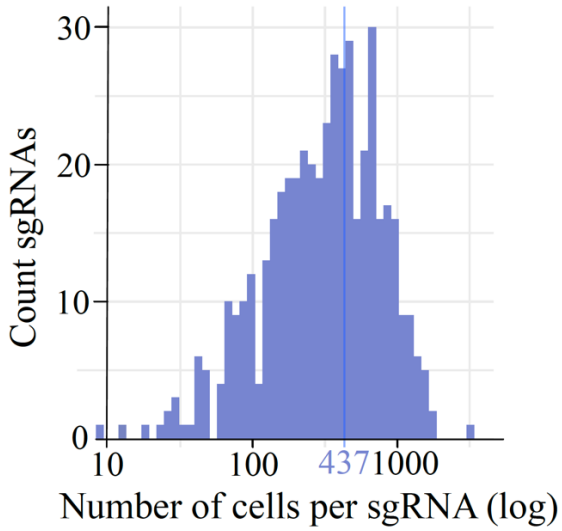



**Figure 4.15– Histogram of sgRNA count in relation to number of cells per sgRNA for the combined dataset**

Histogram of the number of cells across the three transduction replicates expressing a given unique sgRNA from the 475 sgRNA library; the number of cells per sgRNA is depicted in logarithmic scale and the vertical blue line represents the average of 437 cells per sgRNA.

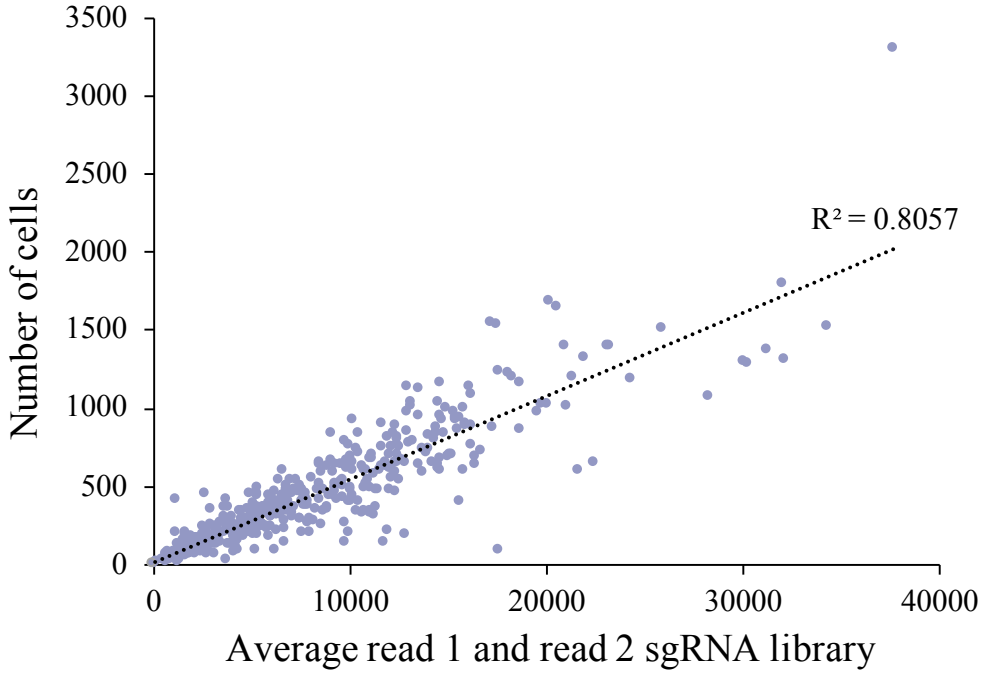



**Figure 4.16 – Scatterplot showing number of cells vs sgRNA representation in the oligo library**

Scatterplot between the number of cells expressing a given sgRNA in the scRNA-seq CRISPRa screen data (y axis) and the normalised representation of the sgRNA in the cloned oligo library before transduction ($log_{10}$ average of read 1 and read 2- x axis). The sgRNA representation in the oligo library and the number of cells expressing that sgRNA are correlated ($R^2$=0.8057).

Both PCA and UMAP (Uniform Manifold Approximation and Projection) are algorithms for high-dimensionality reduction (Pearson 2010; McInnes et al. 2018). Analysis the whole dataset of 203,894 cells expressing a unique sgRNA using both of these methods, on highly variable genes (see Materials and Methods), showed high transcriptional similarity between replicates (Figures 4.17), validating the robustness of the screen. Therefore, cells across the three transduction replicates were merged for downstream analysis to reach the statistical power of at least 367 cells per sgRNA, on average.

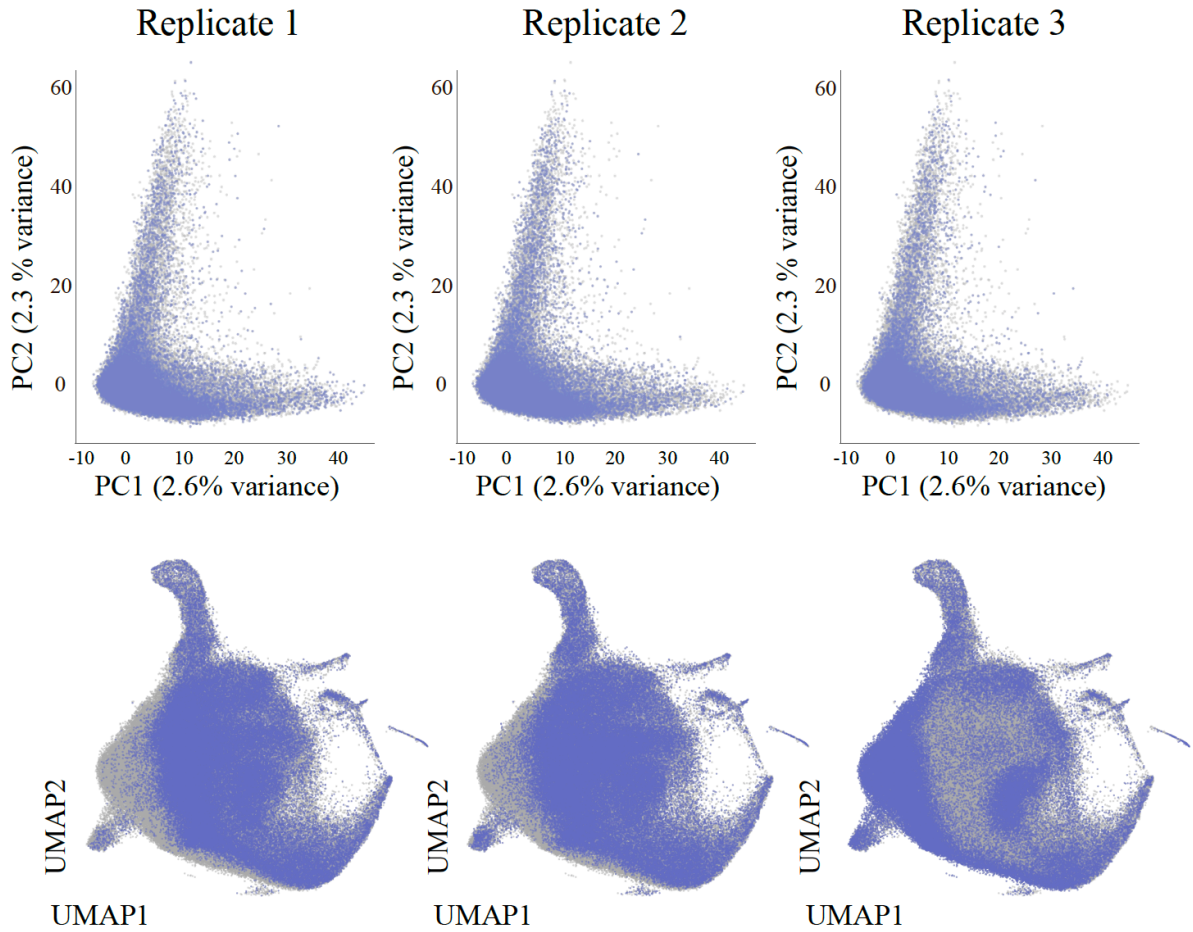



**Figure 4.17– Reproducibility of transduction replicates, analysed by PCA and UMAP**

Principal component analysis (PCA) (top panel) and UMAP (Uniform Manifold Approximation and Projection) (bottom panel) was performed on the whole dataset of 203,894 cells expressing a unique sgRNA. For PCA, the first two principal components (PC1 and PC2) are shown (top panel). PC1 (x axis) explains 2.6% of total variance in the whole dataset whereas PC2 (y axis) explains 2.3% of variance. For UMAP, UMAP factor 1 (UMAP1, x axis) and UMAP factor 2 (UMAP2, y axis) are shown (bottom panel). The cells from each transduction replicate (1-3) contributing to the analysis are shown in each column and highlighted in blue.

### 4.2.4 Target gene activation and CRISPR activation performance

Before I could identify candidate genes whose activation triggered a ZGA-like signature, I first needed to systematically assess the functionality of sgRNAs. In order to do that, the expression of each target gene was analysed in the cells transduced with the corresponding targeting sgRNA in comparison to the basal gene expression in the non-targeting sgRNA control cells (see Materials and Methods). By doing this, activated (cells with higher target gene expression than the basal level in non-targeting sgRNA control cells, FDR<10%, see Materials and Methods) and not-activated (cells with equal or lower target gene expression than the basal level in non-targeting sgRNA control cells, FDR>10%, see Materials and Methods) cells were called for each targeting sgRNA. This method allowed the identification of target genes for which both sgRNAs triggered activation, target genes for which only one of the targeting sgRNAs activated gene expression, and target genes for which none of the sgRNAs upregulated gene expression compared to the non-targeting sgRNA control cells (Figure 4.18). Importantly, the number of activated cells for each sgRNA did not correlate with the total number of cells expressing that sgRNA (Figure 4.19 and Appendix C), indicating that the method established to call activated cells is independent on the total number of cells and that CRISPRa efficiencies at the single-cell level are variable from one sgRNA to another.

Using this method, I considered that a sgRNA had worked effectively when at least 10 transduced cells showed target gene activation. I set this thereshold on the basis that technical drop-outs are common in scRNA-seq data (Qiu et al. 2018) and, therefore, cells classified as not-activated could be the result of biological lack of target gene activation or technical lack of target gene activation detection. Moreover, 10 would be the minimum number of single-cells to be analysed per sgRNA for confident downstream analysis. Using this criterion, a total of 108 sgRNAs were found to be functional (Appendix C). This corresponded to effective target gene activation for 82 out of 230 candidates, 26 of which were activated with both targeting sgRNAs and 56 with only one sgRNA (Figure 4.20 and Appendix C). Given the surprising variability in sgRNA efficiency not only between sgRNAs (Figure 4.19, 4.20) but also between single cells for a given sgRNA (Figure 4.18), I decided to maximise the power of my comprehensive single-cell transcriptomic dataset to investigate in detail the features that drive CRISPRa efficiency.

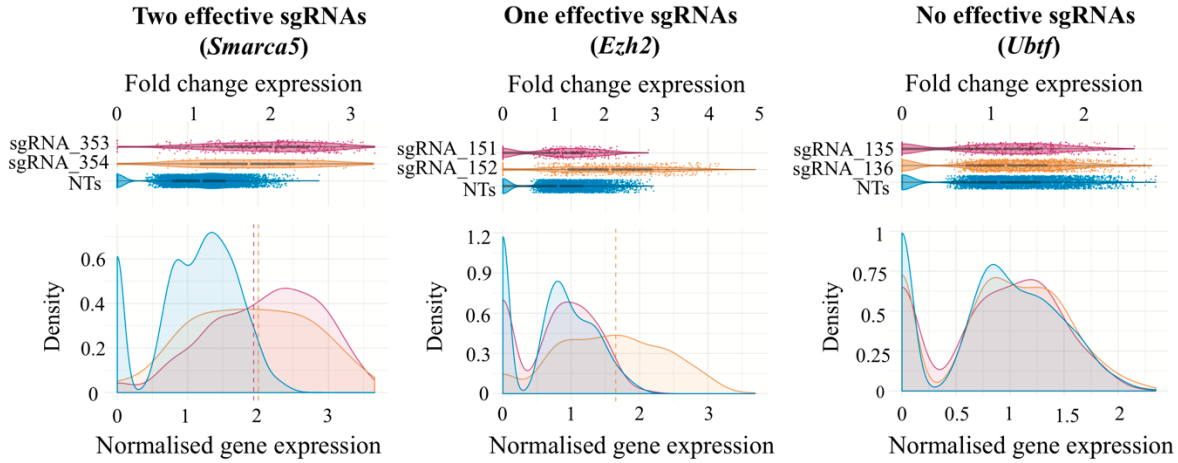



**Figure 4.18– Representatives examples of target gene activation**

Dot-plots (upper panels) and density plots (bottom panels) of representative examples of effective target gene activation with the two targeting sgRNAs (*Smarca5* - left panel), one of the targeting sgRNAs (*Ezh2* - middle panel) or ineffective target gene activation with either of the two targeting sgRNAs (*Ubtf* - right panel). SAM22 mESCs expressing the respective targeting sgRNAs are shown in red and orange and SAM22 mESCs expressing either of the 15 non-targeting sgRNA controls are shown in blue. In the dot-plots, each dot represents a cell plotted as a function of fold change expression to the average expression in the non-targeting sgRNA control cells. In the density plots, normalised gene expression is calculated as $\log_e$(number of UMIs for target gene / sum (number of total UMIs in a cell) *10,000 + 1) and plotted as a function of density. Vertical dashed lines represent an FDR 10% cut-off for activated and not-activated cells within each sgRNA, with cells above the cut-off being activated for the corresponding target gene and cells below the cut-off being not-activated for the target gene. NTs: non-targeting sgRNA controls.

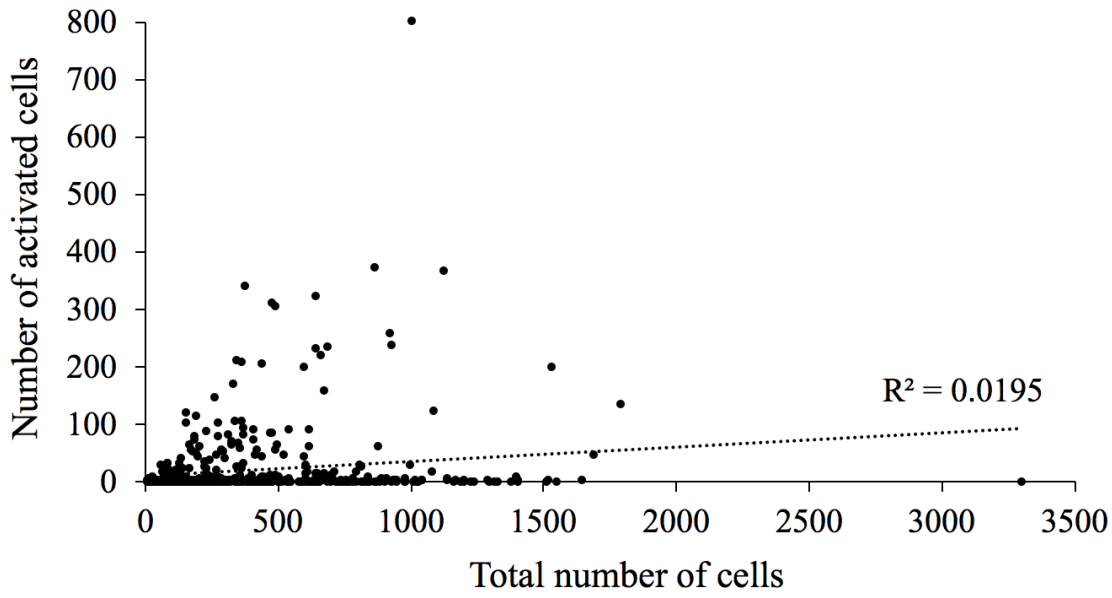



**Figure 4.19– Target gene activation is independent on total number of cells**

Scatterplot between number of activated cells for each sgRNA (y axis) against total number of cells expressing that sgRNA (x axis), showing there is no linear correlation between the two ($R^2 = 0.0195$).

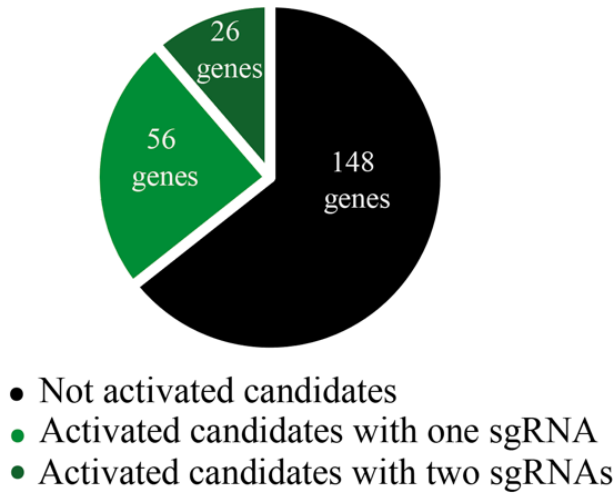



**Figure 4.20– Pie chart of number of candidate genes activated**

Pie chart showing number of candidate genes not activated by either of the two targeting sgRNAs (black), number of candidate genes effectively activated with one of the targeting sgRNAs (light green), and number of candidate genes effectively activated with the two targeting sgRNAs (dark green). A sgRNA was considered effective when at least 10 transduced cells showed target gene activation.

Even though the screen was performed on the clonal SAM22 mESC line, I first asked whether some cells might had lost expression of the stably integrated CRISPRa machinery (dCas9-VP64 and MS2-p65-HSF1) and, if that was the case, whether that explained the differences observed in CRISPRa efficiency. For that, scRNA-seq reads that did not map to the mouse genome assembly were mapped to the integration sequence of dCas9-VP64 and/or MS2-p65-HSF1 plasmids (see Materials and Methods). The two plasmids are identical between themselves in the region immediately upstream of the polyadenylation signal (Figure 3.2), making it impossible to distinguish between the two in 3'-primed scRNA-seq libraries. Furthermore, given the sequence similarity of these two plasmids to the CROP-sgRNA-MS2 plasmid in this same region upstream of the polyadenylation signal, and the limited sequencing read length at the 3' end of transcripts in 10X Genomics scRNA-seq libraries (see Materials and Methods), very few reads mapped confidently to dCas9-VP64 and/or MS2-p65-HSF1 transcripts (Figure 4.21). Therefore, with 3'-primed 10X Genomics scRNA-seq libraries, it is not possible to assess expression of CRISPRa machinery.

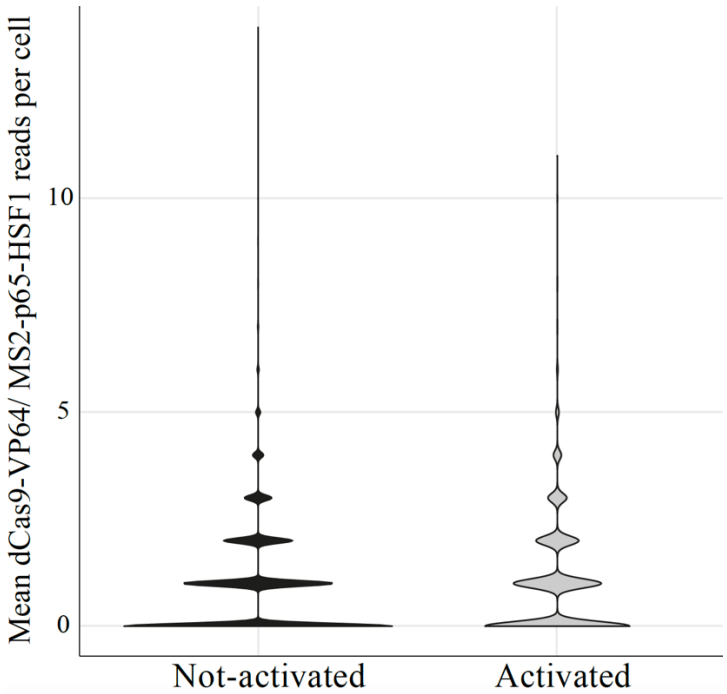



**Figure 4.21– dCas9-VP64 and/or MS2 p65-HSF1 expression in not-activated and activated cells**

Violin plots showing number of dCas9-VP64/MS2-p65-HSF1 reads per cell in not-activated and activated cells for any given sgRNA. Differences between the two groups assessed by Mann-Whitney test were not statistically significant

Before sequencing, the cells transduced with the pooled sgRNA library were selected with puromycin for expression of the sgRNA (Figure 4.10). However, it was possible to think that not-activated cells had lower sgRNA expression due to the random integration nature of lentiviruses, reasoning that not-activated cells could have integrated the sgRNA construct in a genomic site that allowed expression to sufficient levels to survive puromycin selection but not to trigger efficient target gene activation. Using the amplicon sgRNA libraries, I could estimate the number of sgRNA reads per cell. Surprisingly, I found no differences in sgRNA expression between activated and not-activated cells (Figure 4.22), discarding this as a cause for ineffective target gene activation.

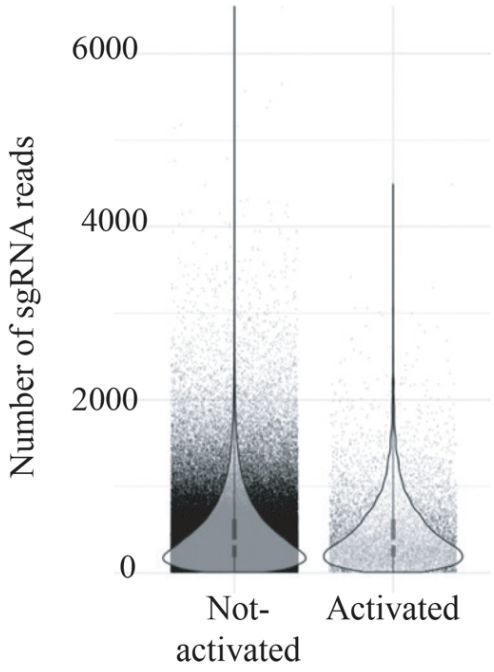



**Figure 4.22– sgRNA expression in not-activated and activated cells**

Dot-plots and violin plots showing number of sgRNA reads in not-activated and activated individual cells for any given sgRNA. sgRNA reads were quantified from the sgRNA amplicon libraries. Differences between the two groups assessed by Mann-Whitney test were not statistically significant.

Next, I hypothesised that features of the sgRNAs themselves are responsible for CRISPRa efficiency, since for 56 out of the 82 effectively activated candidates, only one of the sgRNAs was effective (Figure 4.20). The targeting position of the sgRNA has been shown to be crucial both for CRISPRa and CRISPRi and related to nucleosome occupancy (Konermann et al. 2015; Braun et al. 2016; Horlbeck et al. 2016; Horlbeck, Witkowsky, et al. 2016; Radzisheuskaya et al. 2016). All sgRNAs in my custom library were targeted to the 180 bp region upstream of the TSS of the candidate gene, as this was previously shown to be optimal for CRISPRa (Konermann et al. 2015). Interestingly, I observed that, within this 180 bp window, the targeting position of the sgRNA relative to the TSS was significantly different between effective and ineffective sgRNAs, with effective sgRNAs mapping further away from the TSS than the ineffective ones (Figure 4.23). This difference was more pronounced between effective and ineffective sgRNAs within activated candidates in the dataset (Figure 4.23). These results suggest that the targeting position of the sgRNA was crucial for optimal efficiency, whereas for candidate genes that were not activated with either of the two targeting sgRNAs, other factors might contribute to make those genes less susceptible to activation.

sgRNA base content is also key for CRISPR efficiency and specificity (T. Wang et al. 2014; Radzisheuskaya et al. 2016; Chari et al. 2015; Doench et al. 2016; H. Xu et al. 2015; X. Xu et al. 2017; Graf et al. 2019). While I found no differences in overall GC content between effective and ineffective sgRNAs (Figure 4.24), base-by-base analysis revealed that base composition at positions 8, 15 and 17 (with position 20 being just upstream of PAM) was different for effective and ineffective sgRNAs (Figure 4.25). Positions 8 and 15 were enriched in “C” and position 17 enriched in “G” in effective sgRNAs, whereas ineffective sgRNAs within effectively activated targets were enriched in “G” at position 15 (Figures 4.25, 4.26). Statistical quantification of the proportions of each base at each position in each group, in comparison to the expected proportions, showed that the presence of a “G” at position 15 was predicted for ineffective sgRNAs, and presence of a “C” at position 8 was also close to be statistically significant (p-value = 0.07) for effective sgRNAs (Figure 4.26). Similar to the effect observed for the sgRNA targeting distance to the TSS (Figure 4.23), these differences were overall more pronounced between effective and ineffective sgRNAs within activated

candidates in the dataset, suggesting again that other factors might play a role in the susceptibility of a gene to become activated despite the effectiveness of the sgRNA.

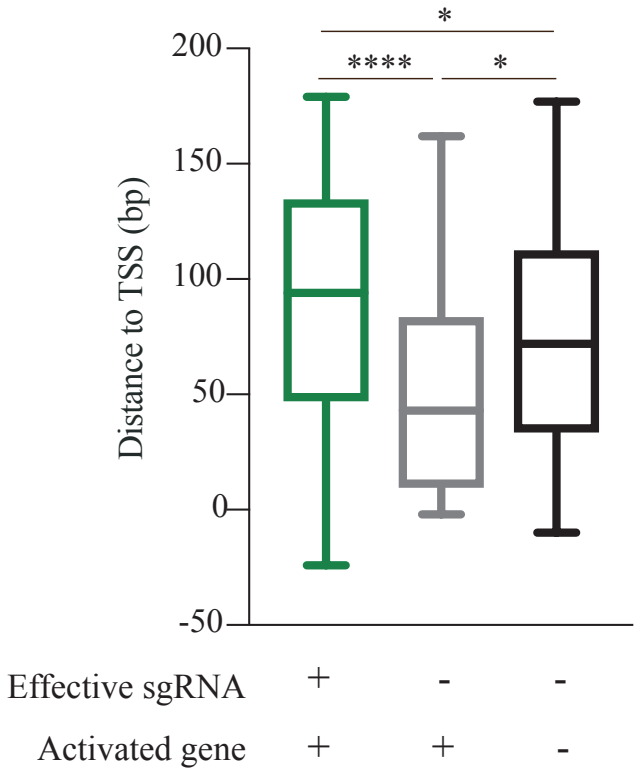



**Figure 4.23– sgRNA targeting position for effective and ineffective sgRNAs**

Box-whisker plots showing mapping distance in base-pairs (bp) from the 3' end of the sgRNA (base proximal to PAM) to the TSS, for effective sgRNAs (green) and ineffective sgRNAs (grey and black); ineffective sgRNAs are sub-divided into two categories: those targeting activated candidate genes in the dataset by the alternative sgRNA (grey) and those targeting not-activated candidate genes in the dataset by either of the two sgRNAs (black). A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Statistically significant differences between groups were assessed by one-way ANOVA test: ****: p-value < 0.0001, * p-value <0.5.

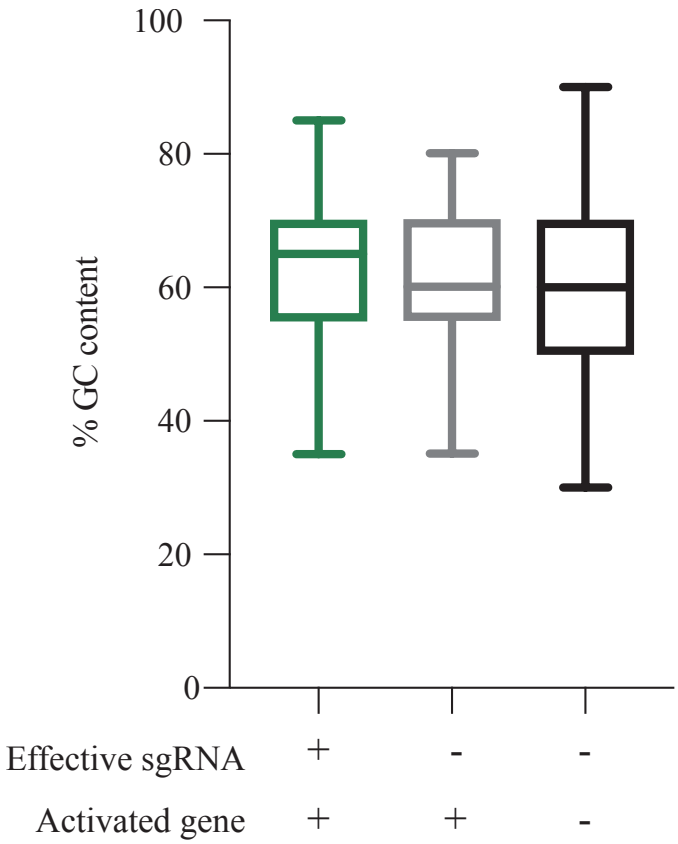



**Figure 4.24– Percentage of GC content in effective and ineffective sgRNAs**

Box-whisker plots showing percentage of GC content in effective sgRNAs (green) and ineffective sgRNAs (grey and black); ineffective sgRNAs are sub-divided into two categories: those targeting activated candidate genes in the dataset by the alternative sgRNA (grey) and those targeting not-activated candidate genes in the dataset by either of the two sgRNAs (black). A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between groups assessed by one-way ANOVA test were not statistically significant (ns).

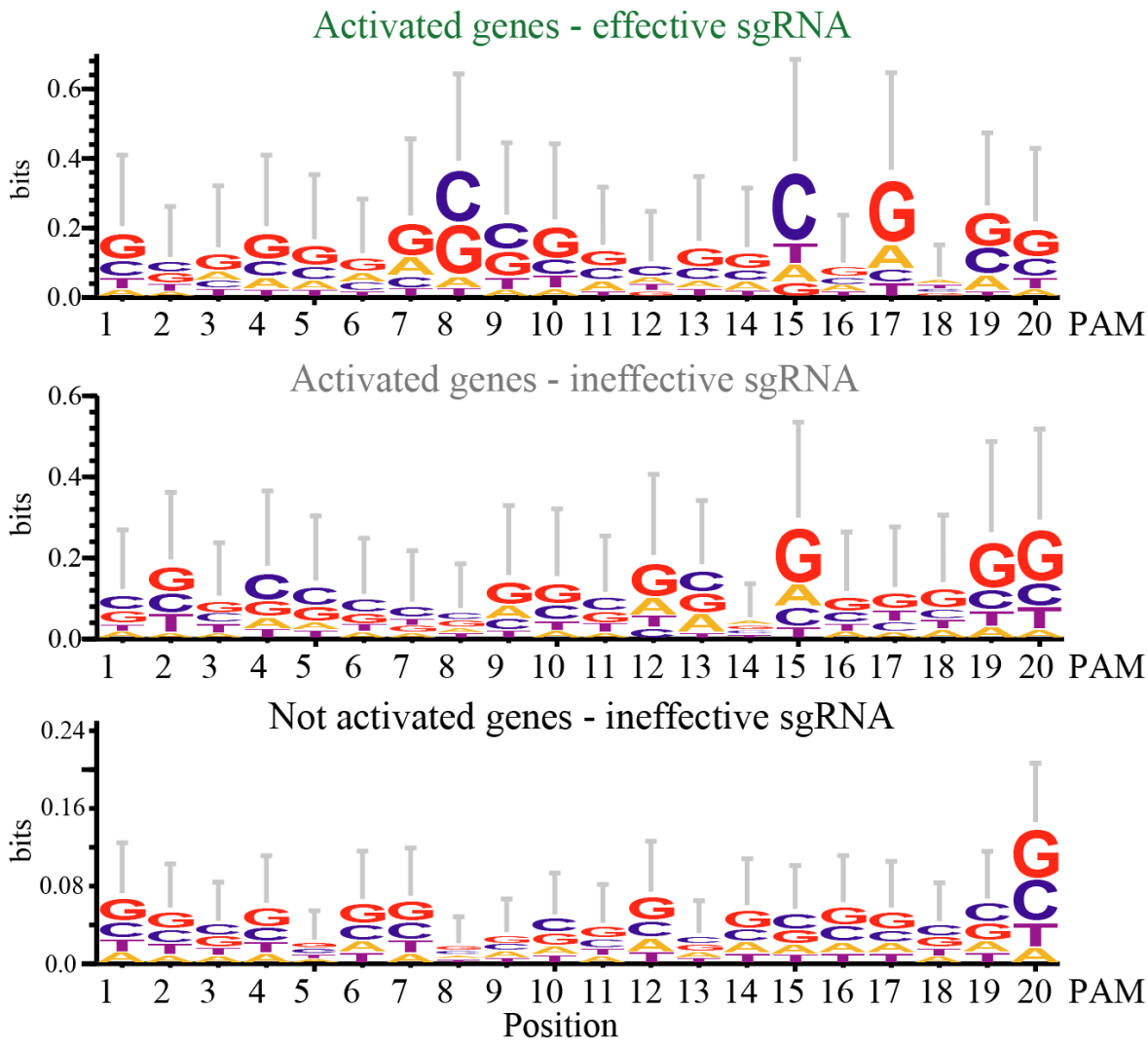



**Figure 4.25– Base content at each position of effective and ineffective sgRNAs**

Logo representation of sgRNA base content at each position of effective sgRNAs (top panel - green) and ineffective sgRNAs (middle and bottom panels - grey and black); ineffective sgRNAs are sub-divided into two categories: those targeting activated candidate genes in the dataset by the alternative sgRNA (middle panel - grey) and those targeting not-activated candidate genes in the dataset by either of the two sgRNAs (bottom panel - black). Position 1 is the furthest away from the PAM region and base 20 is just upstream of it. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation.

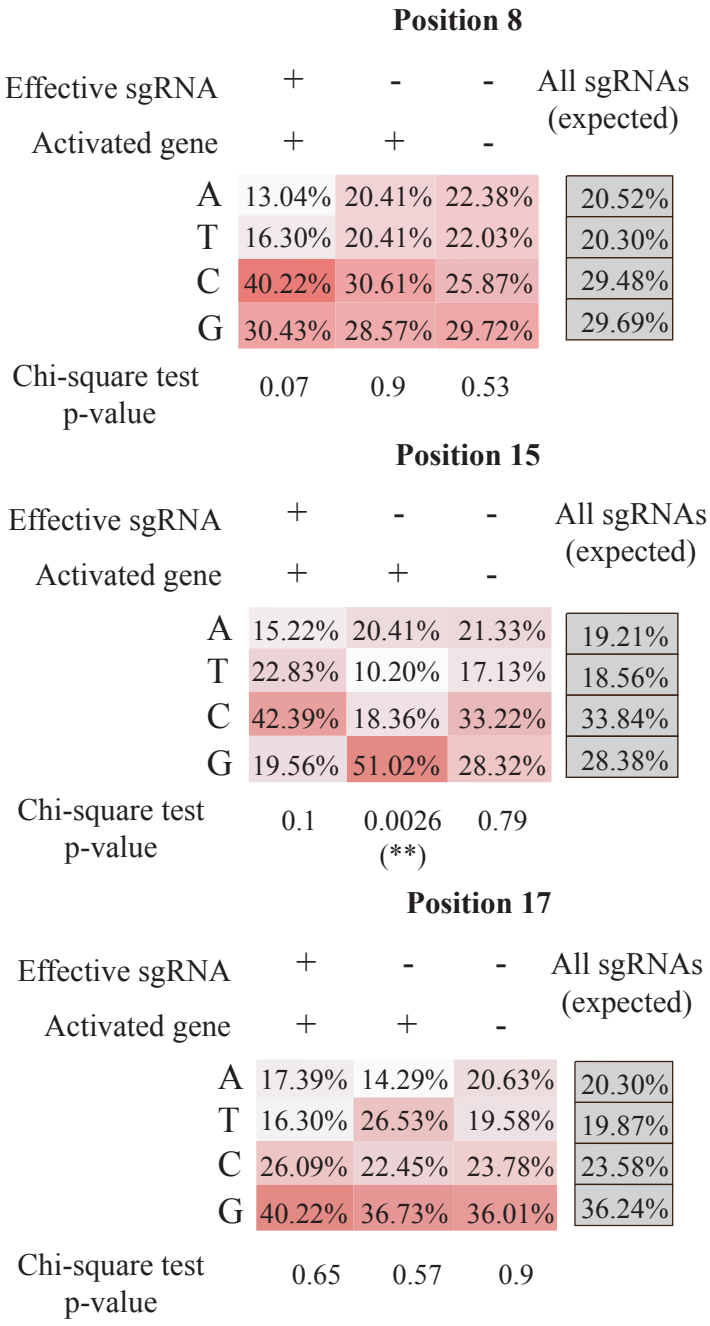



**Figure 4.26– Quantification of base content at each position of effective and ineffective sgRNAs**

Heatmaps showing percentage of each base in sgRNA positions 8 (top panel), 15 (middle panel) and 17 (bottom panel) for effective sgRNAs (left columns) and ineffective sgRNAs (middle and right columns), as well as the percentage of each base in the total 460 targeting sgRNAs in the library (grey). Ineffective sgRNAs are sub-divided into two categories: those targeting activated candidate genes in the dataset by the alternative sgRNA (middle columns) and those targeting not-activated candidate genes in the dataset by either of the two sgRNAs (right columns). The p-value of chi-square tests for the proportions of each base in each sgRNA category and each position in comparison to the expected proportion is shown. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation.

Therefore, I next grouped the targeted genes into "activated" (with one or both sgRNAs) and "not-activated" genes and investigated the differences, if any, in their basal expression as well as multiple epigenetic features. The CRISPRa SAM system has been shown to work better for lowly expressed target genes, with fold changes achieved being anti-correlated with basal gene expression (Konermann et al. 2015; Chavez et al. 2016). However, I found no differences in basal gene expression between activated and not-activated genes in my screen when I looked at their normalised expression in untransduced SAM22 mESCs bulk RNA-sequencing libraries (Figure 4.27) or in E14 mESCs bulk RNA-sequencing libraries from multiple datasets and studies (Barisic et al. 2019, von Meyenn et al. 2016, this dissertation) (data not shown).

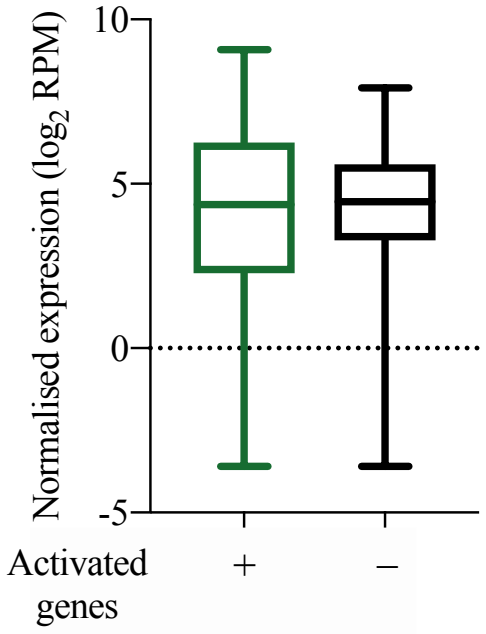



**Figure 4.27– Basal expression of activated and not-activated genes in SAM22 mESCs by bulk RNA-sequencing**

Box-whisker plots showing basal normalised expression in $\log_2$ reads per million (RPM) in SAM22 mESCs of effectively activated (green) and not-activated (black) gene candidates in the screen, analysed by bulk RNA-sequencing. A gene was considered activated when at least one of the targeting sgRNAs lead to efficient target gene activation. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between the two groups assessed by Mann-Whitney test were not statistically significant.

Since gene expression can be heterogeneous in mESCs (reviewed in Tanaka 2009), I also analysed basal expression levels of activated and not-activated genes in cells transduced with non-targeting sgRNAs in the screen 10X Genomics scRNA-seq libraries as well as in

untransduced E14 mESCs analysed by either 10X Genomics or Smart-Seq2 scRNA-seq. E14 10X Genomics scRNA-seq libraries were described in chapter 3 and E14 Smart-Seq2 scRNA-seq libraries were published in Rulands et al. 2018 (see Materials and Methods). Unexpectedly, in all three cases, I observed that the basal level of gene expression was higher for effectively activated gene targets (Figure 4.28).

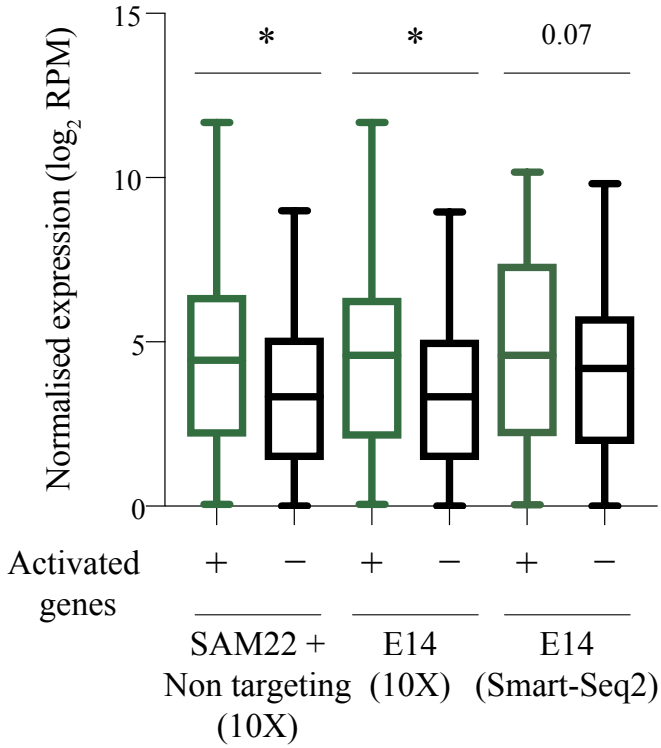



**Figure 4.28– Basal expression of activated and not-activated genes in scRNA-seq libraries**

Box-whisker plots showing basal normalised expression in $\log_2$ reads per million (RPM) of effectively activated (green) and not-activated (black) gene candidates in the screen, in SAM22 mESCs transduced with non-targeting sgRNAs and analysed by 10X Genomics scRNA-seq and in E14 mESCs analysed by either 10X Genomics scRNA-seq or by Smart-Seq2 scRNA-seq (E14 Smart-Seq2 data analysed from Rulands et al. 2018). A gene was considered activated when at least one of the targeting sgRNAs lead to efficient target gene activation. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between activated and not-activated genes for each of the three cell types and/or platforms of sequencing were assessed by Mann-Whitney test: * p-value <0.05, p-value = 0.07 for E14 mESCs analysed by Smart-Seq2.

Observing differences in basal gene expression between these two groups by scRNA-seq but not by bulk RNA-sequencing suggested heterogeneity at the single-cell level in the expression of these genes. However, I observed no differences in the variability of basal expression between activated and not-activated genes (Figure 4.29) (see Materials and Methods).

Therefore, it is difficult to conclude whether the higher basal gene expression observed for activated genes in scRNA-seq libraries makes these genes more susceptible to activation by CRISPRa SAM in mESCs or whether the differences observed are a technical artifact of scRNA-seq by which these genes are generally detected at higher levels.

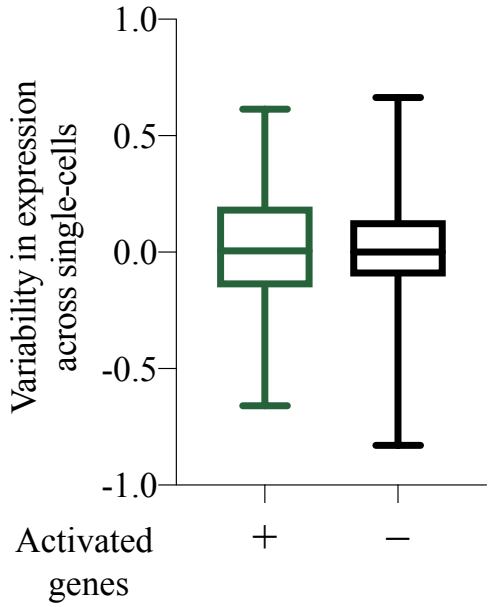



**Figure 4.29– Basal expression variability of activated and not-activated genes by scRNA-seq**

Box-whisker plots showing basal variability in expression across single-cells (see Materials and Methods) of effectively activated (green) and not-activated (black) gene candidates in the screen, in E14 mESCs analysed by Smart-Seq2 scRNA-seq (data analysed from Rulands et al. 2018). A gene was considered activated when at least one of the targeting sgRNAs lead to efficient target gene activation. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between the two groups assessed by Mann-Whitney test were not statistically significant.

To distinguish between these two possibilities, I next used published sequencing datasets of mESCs to analysed a wide range of basal epigenetic features of these genes that might influence their expression levels. First, I analysed ATAC-seq (King et al. 2018; Dongwei Li et al. 2017) and DNase-seq (ENCODE) datasets and determined that the accessibility of the chromatin at basal levels was not different in gene promoters of activated and not-activated genes (Figure 4.30).

The histone post-translational modifications H3K4me3 and H3K27ac are generally associated with transcriptional activity and open chromatin, respectively (reviewed in Kouzarides 2007) (Figure 1.6). However, and consistent with the results observed by ATAC-seq and DNase-seq, no differences were detected on average in these two marks at the promoters of activated and not-activated genes by ChIP-seq (Figure 4.31) (data analysed from Hernandez et al. 2018).

Conversely, H3K27me3 and H3K9me3 are both marks associated with inactive chromatin and repressive gene expression (reviewed in Kouzarides 2007) (Figure 1.6). Overall, the candidate genes included in the screen are depleted of H3K27me3 and H3K9me3 at their promoters in mESCs (Figure 4.32), consistent with their expression levels (Figure 4.3) and regardless of their activation state in the screen (Figure 4.32) (H3K27me3 ChIP-seq data analysed from Hernandez et al. 2018 and H3K9me3 ChIP-seq data analysed from Bilodeau et al. 2009).

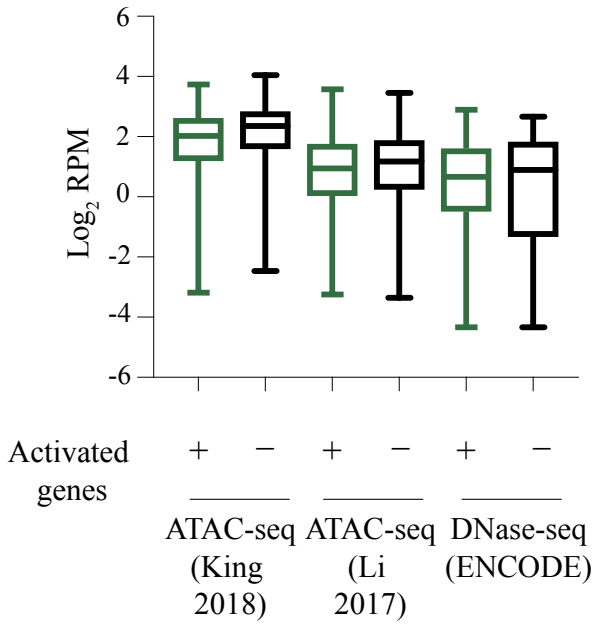



**Figure 4.30– Basal chromatin accessibility in the promoters of activated and not-activated genes**

Box-whisker plots showing basal chromatin accessibility of gene promoters of effectively activated (green) and not-activated (black) gene candidates in the screen, measured as $log_2$ reads per million (RPM) in the region 1 kilobase upstream of the gene TSS in mESC ATAC-seq (King et al. 2018, Dongwei Li et al. 2017) and DNase-seq (ENCODE) libraries. A gene was considered activated when at least one of the targeting sgRNAs lead to efficient target gene activation. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between activated and not-activated genes in each dataset were assessed by Mann-Whitney test and were not statistically significant.

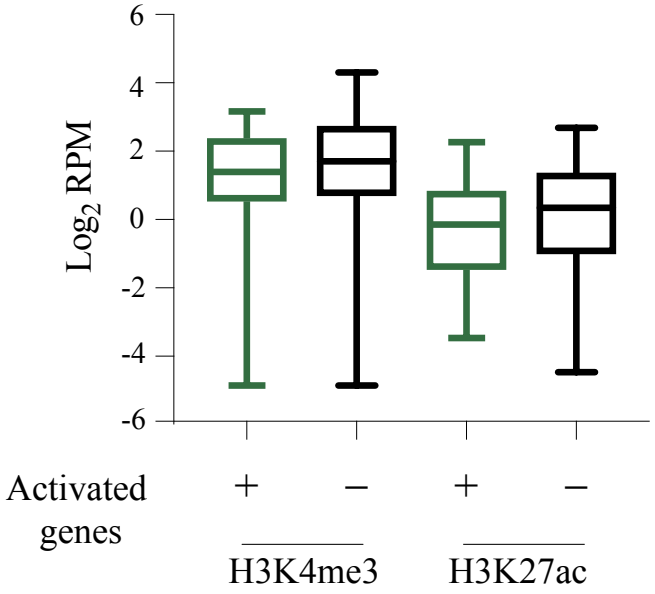



**Figure 4.31– Basal H3K4me3 and H3K27ac in the promoters of activated and not-activated genes**

Box-whisker plots showing basal H3K4me3 and H3K27me3 at gene promoters of effectively activated (green) and not-activated (black) gene candidates in the screen, measured as $log_2$ reads per million (RPM) in the region 1 kilobase upstream of the gene TSS in H3K4me3 and H3K27ac mESC ChIP-seq libraries (data analysed from Hernandez et al. 2018). A gene was considered activated when at least one of the targeting sgRNAs lead to efficient target gene activation. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between activated and not-activated genes for each histone mark were assessed by Mann-Whitney test and were not statistically significant.

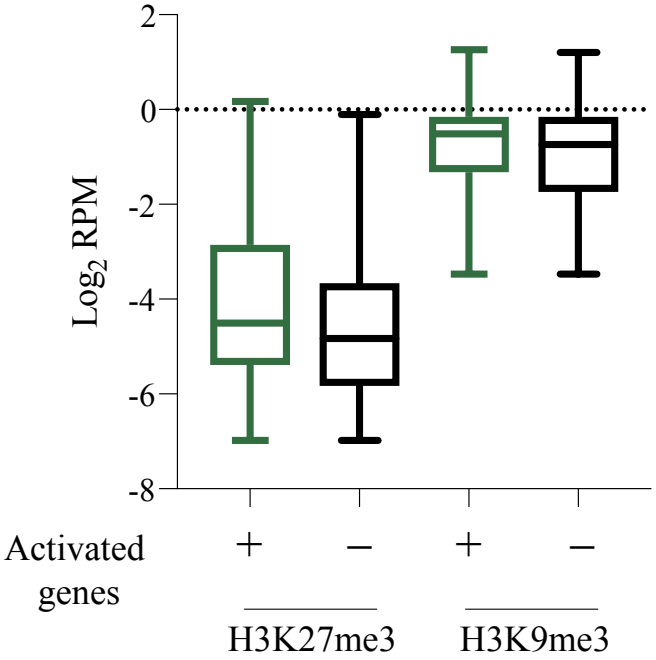



**Figure 4.32– Basal H3K27me3 and H3K9me3 in the promoters of activated and not-activated genes**

Box-whisker plots showing basal H3K27me3 and H3K9me3 at gene promoters of effectively activated (green) and not-activated (black) gene candidates in the screen, measured as $log_2$ reads per million (RPM) in the region 1 kilobase upstream of the gene TSS in H3K27me3 mESC ChIP-seq libraries (analysed from Hernandez et al. 2018) and H3K9me3 mESC ChIP-seq libraries (analysed from Bilodeau et al. 2009). A dashed line at y = 0 indicates that the promoters analysed are largely depleted of H3K27me3 and H3K9me3 marks in mESCs. A gene was considered activated when at least one of the targeting sgRNAs lead to efficient target gene activation. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between activated and not-activated genes for each histone mark were assessed by Mann-Whitney test and were not statistically significant.

DNA methylation at gene promoters has long been associated to transcriptional repression (Razin & Riggs 1980) via different mechanisms such as recruitment of histone post-translational modifications or by preventing the binding of transcription factors (reviewed in Klose & Bird 2006) (Figure 1.6). Therefore, I analysed the DNA methylation state of the gene promoters of activated and not-activated genes in whole-genome bisulfite sequencing libraries (Ficz et al. 2013, Habibi et al. 2013, Milagre et al. 2017, Berrens et al. 2017) and found overall low levels of DNA methylation that did not differ between the two groups (Figure 4.33).

Lastly, H3K36me3, which is often found in gene bodies of actively transcribed genes (reviewed in Kouzarides 2007) (Figure 1.6), was overall present at basal levels in all screen candidates and its average occurrence was not different between the groups of activated and not activated genes (Figure 4.34) (data analysed from Neri et al. 2017). Altogether, these observations point to an overall similar basal epigenetic and transcriptional state between the group of activated and not-activated genes in the screen.

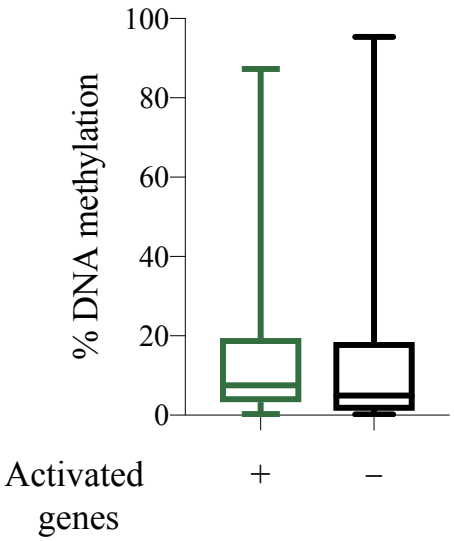



**Figure 4.33– Percentage of basal DNA methylation in the promoters of activated and not-activated genes**

Box-whisker plots showing percentage of basal CpG methylation at gene promoters of effectively activated (green) and not-activated (black) gene candidates in the screen, analysed by whole-genome bisulfite sequencing (Ficz et al. 2013, Habibi et al. 2013, Milagre et al. 2017, Berrens et al. 2017). Whole-genome bisulfite libraries from wild-type mESCs from the four studies were merged to increase sequencing coverage. Gene promoters are defined as the region 1 kilobase upstream of the gene TSS. Percent DNA methylation was calculated as the percent of methylated CpGs over total CpGs in the region of interest. A gene promoter was considered for analysis if it had at least 10 CpGs covered by at least 10 reads. A gene was considered activated when at least one of the targeting sgRNAs lead to efficient target gene activation. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between activated and not-activated genes were assessed by Mann-Whitney test and were not statistically significant.

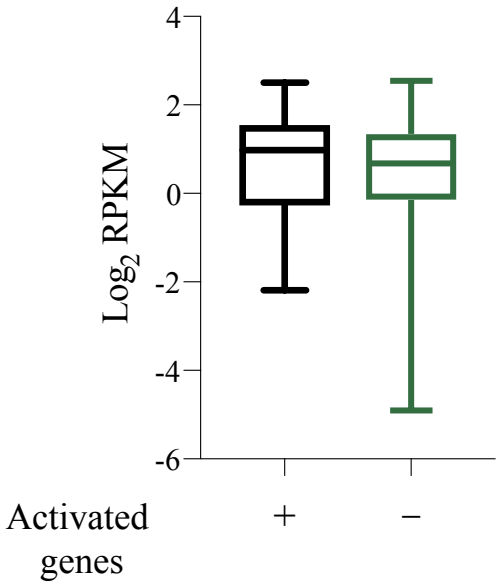



**Figure 4.34– Basal H3K36me3 in the gene bodies of activated and not-activated genes**

Box-whisker plots showing basal H3K36me3 in gene bodies of effectively activated (green) and not-activated (black) gene candidates in the screen, measured as $log_2$ reads per kilobase per million (RPKM) over gene bodies in H3K36me3 mESC ChIP-seq libraries (data analysed from Neri et al. 2017). A gene was considered activated when at least one of the targeting sgRNAs lead to efficient target gene activation. A sgRNA was considered effective when at least 10 transduced cells showed target gene activation. Differences between activated and not-activated genes were assessed by Mann-Whitney test and were not statistically significant.

Moreover, I also performed transcription factor motif analysis on the promoters of activated genes on the one hand and the promoters of not-activated genes on the other hand (see Materials and Methods). However, I did not find motifs differentially enriched between the groups (data not shown), suggeting that there are no transcription factors that specifically mediate CRISPRa SAM to contribute to its efficiency in the target genes tested.

To sum up, the sgRNA targeting position within the 180 bp window upstream of the gene TSS determined transcriptional upregulation mediated by CRISPRa SAM in my screen (Figure 4.23). However, within the genes that were not effectively activated by either sgRNA, there might be multiple factors, including the targeting position of the sgRNAs, but also others such as basal gene expression or epigenetic state, that individually cannot explain why this group of genes was not activated.

As also noted earlier, not all cells transduced with a given effective sgRNA showed target gene activation (Figures 4.18, 4.19) and this was not due to differences in sgRNA expression (Figure 4.22). Therefore, I next sought to investigate the differences, if any, in the epigenetic state of activated and not-activated cells, reasoning that the epigenetic heterogeneity present in mESC cultures (Rulands et al. 2018; Luo et al. 2018; Clark et al. 2018) might influence the ability of a cell to activate the target gene by CRISPRa, i.e., a cell with a closed chromatin state in the promoter of the target gene might be less susceptible to activation than a cell with a wider open chromatin in the target gene promoter. To address this question, I used scNMT-seq (single-cell NOMe, methylation and transcription-sequencing) (Clark et al. 2018), a technique that allows profiling chromatin accessibility, DNA methylation and transcription from the same single cell. Using this method, I could assess target gene activation by CRISPRa SAM at the transcriptional level and correlate it to the chromatin accessibility and DNA methylation state of the target gene promoter in each individual cell. Due to the cost and throughput limitations of this technique, it was not possible to profile all the cells in the screen targeted with the 475 sgRNA library. For this reason, I selected one effective sgRNA, Dppa3 sgRNA 1 (or sgRNA 379, Appendix C), previously shown to activate *Dppa3* expression both at the mRNA (Figure 3.29) and protein levels (Figure 3.30), but at different efficiencies in individual cells (Figure 4.35).

After transducing SAM22 mESCs with the Dppa3 sgRNA or with a non-targeting sgRNA control, I sorted 44 single cells from each condition into a GpC methyltransferase-containing mix, which was used to label open chromatin. Subsequently, the RNA and DNA were physically separated using oligo(d)T-conjugated beads. Then, the RNA was processed to obtain the transcriptome information, and the DNA was bisulfite-treated and DNA libraries processed to obtain chromatin accessibility (GC methylation) and CpG DNA methylation information (Clark et al. 2018) (see Materials and Methods).

After quality control of the scRNA-seq libraries, 38 cells transduced with Dppa3 sgRNA and 21 cells transduced with non-targeting sgRNA control were kept for analysis (see Materials and Methods). I first corroborated efficient target gene activation showing that *Dppa3* was the highest upregulated gene in targeted cells compared to non-targeting sgRNA control cells (Figure 4.36). Next, I analysed the level of *Dppa3* expression at the single-cell level to determine whether the expression pattern recapitulated what I observed in the 10X Genomics

scRNA-seq libraries, with targeted cells ranging from low to high activation (Figure 4.35). However, the expression of *Dppa3* was less dynamic between single-cells after *Dppa3* targeting in this experiment, with 37 out of 38 *Dppa3*-targeted cells showing levels of *Dppa3* expression higher than the non-targeting sgRNA control (Figure 4.37). This was probably due to the lower number of cells analysed compared to the 10X Genomics scRNA-seq dataset.

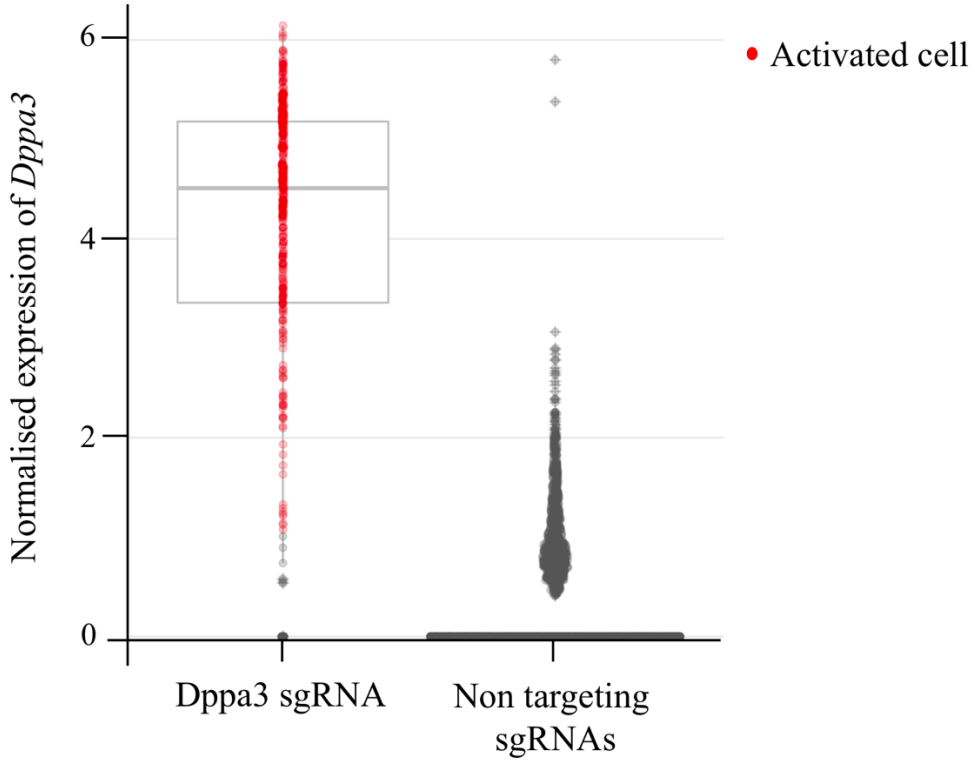



**Figure 4.35– *Dppa3* expression by CRISPRa in the screen 10X Genomics scRNA-seq data**

Box-whisker and dot-plot showing normalised *Dppa3* expression ($\log_e$(number of *Dppa3* UMIs / sum (number of total UMI in a cell) *10,000 + 1)) in SAM22 mESCs transduced with Dppa3 sgRNA 379 or with either of the 15 non-targeting sgRNA controls (see Appendix C), analysed by 10X Genomics 3' scRNA-seq in the CRISPRa screen dataset. Each dot represents a single cell. Red dots represent cells with activated *Dppa3* expression. A cell was considered activated when *Dppa3* expression levels were above a 10% FDR cut-off compared to non-targeting sgRNA controls.

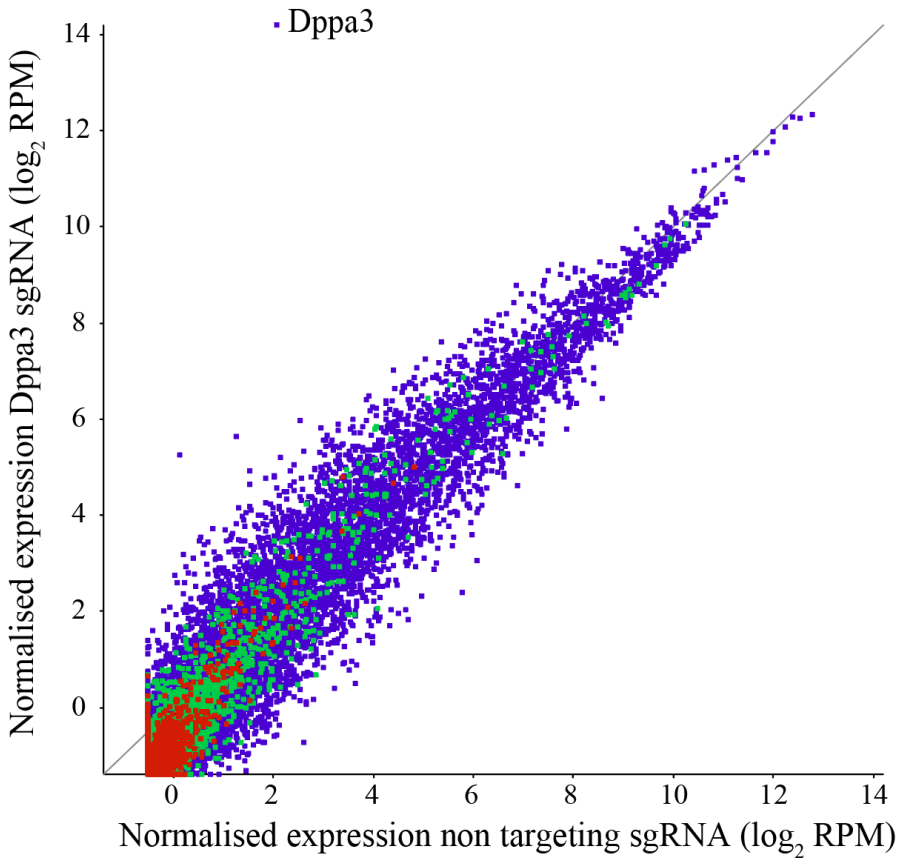



**Figure 4.36– Pseudo-bulked transcriptome upon *Dppa3* CRISPRa, analysed by scNMT-seq**

Scatterplot showing normalised gene expression in $log_2$ reads per million (RPM) of pseudo-bulked SAM22 mESCs transduced with Dppa3 sgRNA 379 (y axis) or with non-targeting sgRNA control 462 (x axis) (see Appendix C), analysed by scNMT-seq, highlighting *Dppa3* transcript.

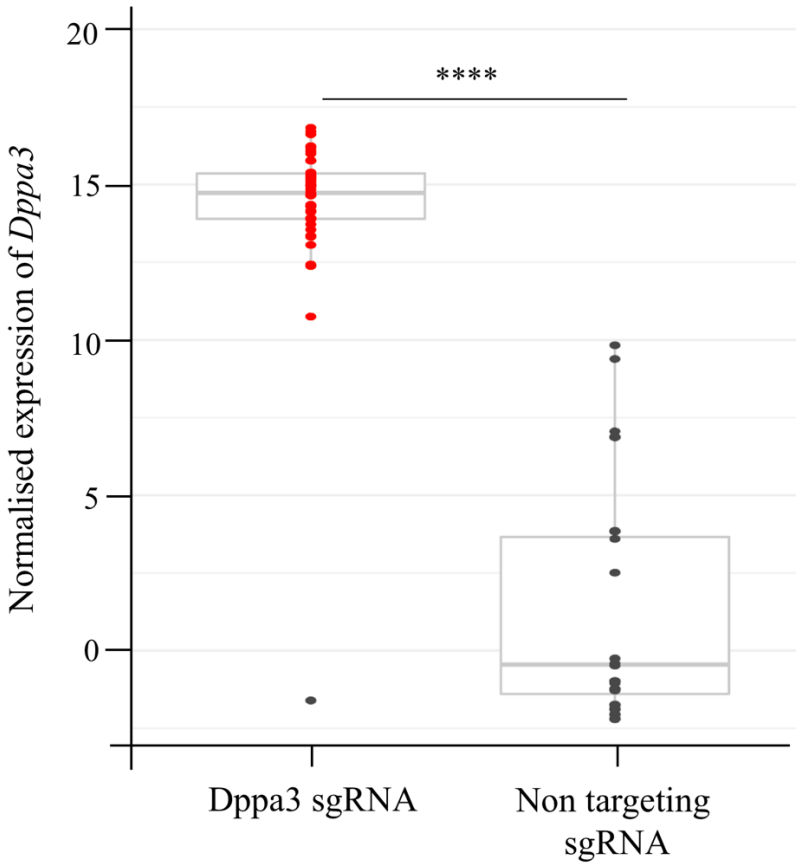



**Figure 4.37– *Dppa3* expression by CRISPRa, analysed by scNMT-seq**

Box-whisker and dot-plot showing normalised *Dppa3* expression in $\log_2$ reads per million (RPM) in SAM22 mESCs transduced with Dppa3 sgRNA 379 or with non-targeting sgRNA control 462 (grey) (see Appendix C), analysed by scNMT-seq. Each dot represents a single cell. Red dots represent cells with higher *Dppa3* expression than cells transduced with non-targeting sgRNA controls. Differences between the two groups were statistically significant (Mann-Whitney test: p-value < $4.89 \times 10^{-13}$).

Next, I analysed global chromatin accessibility after *Dppa3* CRISPRa. 27 cells transduced with Dppa3 sgRNA and 31 cells transduced with non-targeting sgRNA passed quality control for accessibility (see Materials and Methods). At a global level, *Dppa3* activation induced a slight increase in chromatin accessibility (Figure 4.38). However, I was interested to see whether cells with higher target gene activation also had an increased chromatin accessibility in the promoter of the target gene. For that, I analysed the correlation between *Dppa3* expression and *Dppa3* promoter accessibility. Unfortunately, only 11 cells transduced with Dppa3 sgRNA and 10 cells transduced with non-targeting sgRNA passed quality control for both accessibility at the *Dppa3* promoter and scRNA-seq (see Materials and Methods), reducing significantly the number of cells that could be used for this analysis. Nevertheless, using these few cells, there

was a clear difference in *Dppa3* promoter accessibility between non-targeting sgRNA control cells and *Dppa3*-targeted cells (Figure 4.39). This could reflect the observed global changes in accessibility upon *Dppa3* CRISPRa (Figure 4.38) or, alternatively, a specific opening of the promoter as a result of the recruitment of CRISPRa SAM machinery which, consequently, allows transcriptional activation of the target gene. However, there was no linear relationship between *Dppa3* transcriptional activation by CRISPRa and target promoter accessibility within the *Dppa3*-targeted cells (Figure 4.39). Given the limited range of *Dppa3* expression in cells transduced with Dppa3 sgRNA (Figure 4.37 and 4.39), it was not possible to conclude with this experiment whether the CRISPRa efficiency achieved at the single-cell level within cells expressing the targeting sgRNA correlates with target promoter accessibility.

Next, I assessed the global changes in DNA methylation in the 27 cells transduced with Dppa3 sgRNA and 32 cells transduced with non-targeting sgRNA that passed quality control (see Materials and Methods). *Dppa3* CRISPRa induced a marked decrease in global DNA methylation (Figure 4.40). This is consistent with the observed increase in global chromatin accessibility (Figure 4.38) and with literature showing that DPPA3 prevents *de novo* DNA methylation and regulates the machinery for maintenance methylation (Y. Li et al. 2018; Mulholland et al. 2018).

Similar to the analysis done for chromatin accessibility, I next aimed to find a correlation between DNA methylation in the *Dppa3* promoter and *Dppa3* expression after CRISPRa. However, due to the sparsity of the data, few cells passed quality controls for both expression and DNA methylation data (see Materials and Methods), and the trend observed rather reflected the global DNA methylation changes induced after *Dppa3* CRISPRa than a correlation between target gene overexpression and promoter methylation (Figure 4.41).

Altogether, this scNMT-seq experiment did not provide the answers sought due to the sparsity of the data that makes the method unsuitable to assess chromatin accessibility and DNA methylation changes in small genomic regions such as a single specific gene promoter.

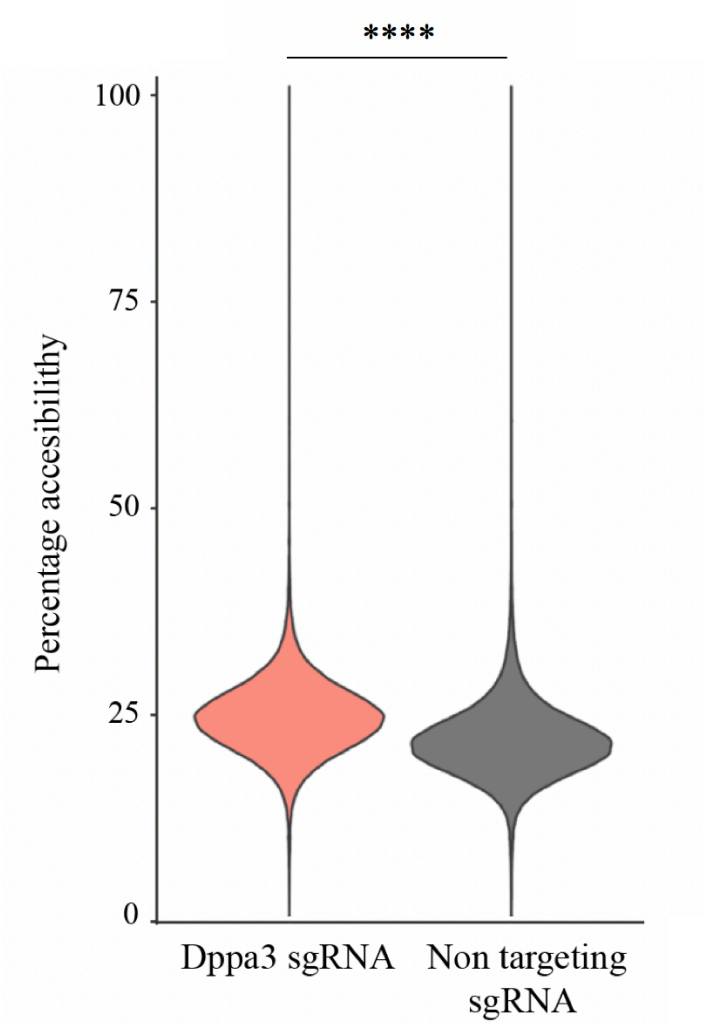



**Figure 4.38– Global chromatin accessibility after *Dppa3* CRISPRa, analysed by scNMT-seq**

Violin plots showing percentage of global chromatin accessibility in SAM22 mESCs transduced with Dppa3 sgRNA 379 (light red) or with non-targeting sgRNA control 462 (grey) (see Appendix C), analysed by scNMT-seq. Percentage accessibility was calculated as the percentage of methylated GpCs over total GpCs in 10 kilobases windows, with a minimum of 10 observations, and the mean accessibility per window across all cells is shown. Differences between the two groups were statistically significant (Mann-Whitney test: p-value < $2.2 \times 10^{-16}$).

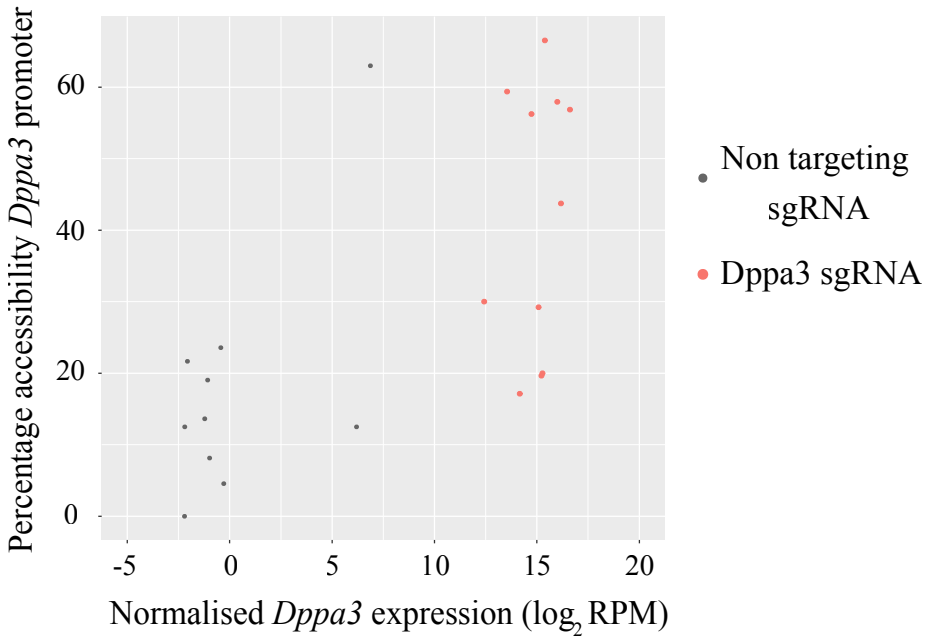



**Figure 4.39– Correlation between *Dppa3* transcriptional activation by CRISPRa and *Dppa3* promoter accessibility**

Scatterplot between normalised expression of *Dppa3* in $\log_2$ reads per million (RPM) (x axis) and percentage chromatin accessibility in the *Dppa3* promoter (y axis) in SAM22 mESCs transduced with Dppa3 sgRNA 379 (light red) or with non-targeting sgRNA control 462 (grey) (see Appendix C), analysed by scNMT-seq. Each dot represents a single cell. Percentage chromatin accessibility in the *Dppa3* promoter was calculated as the percentage of methylated GpCs over total GpCs in cells with a minimum of 10 observations in the region 1.5 kilobases upstream and 0.5 kilobases downstream of the *Dppa3* TSS.

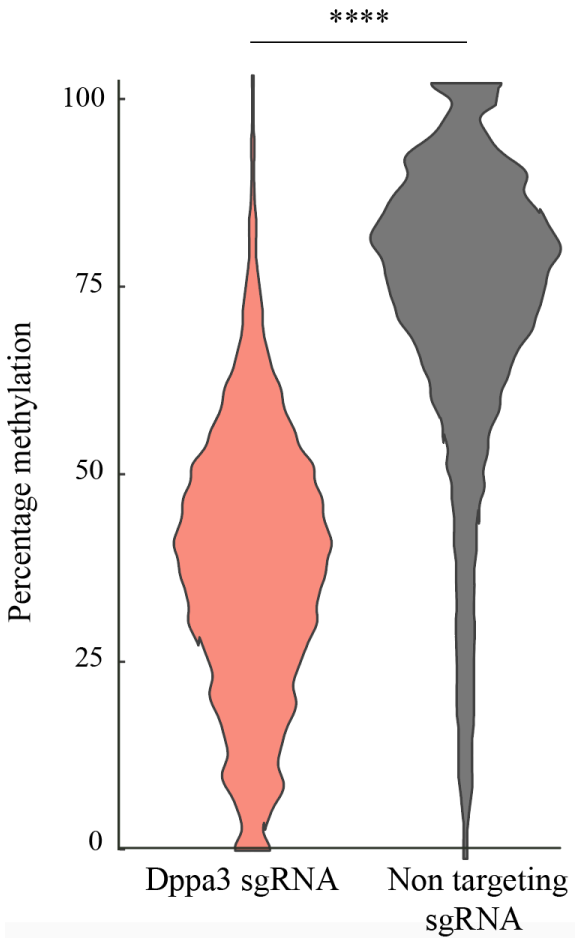



**Figure 4.40– Global DNA methylation after *Dppa3* CRISPRa, analysed by scNMT-seq**

Violin plots showing percentage of global CpG methylation in SAM22 mESCs transduced with Dppa3 sgRNA 379 (light red) or with non-targeting sgRNA control 462 (grey) (see Appendix C), analysed by scNMT-seq. Percentage DNA methylation was calculated as the percent of methylated CpGs over total CpGs in 10 kilobases windows with a minimum of 10 observations, and the mean methylation per window across all cells is shown. Differences between the two groups were statistically significant (Mann-Whitney test: p-value < $2.2 \times 10^{-16}$).

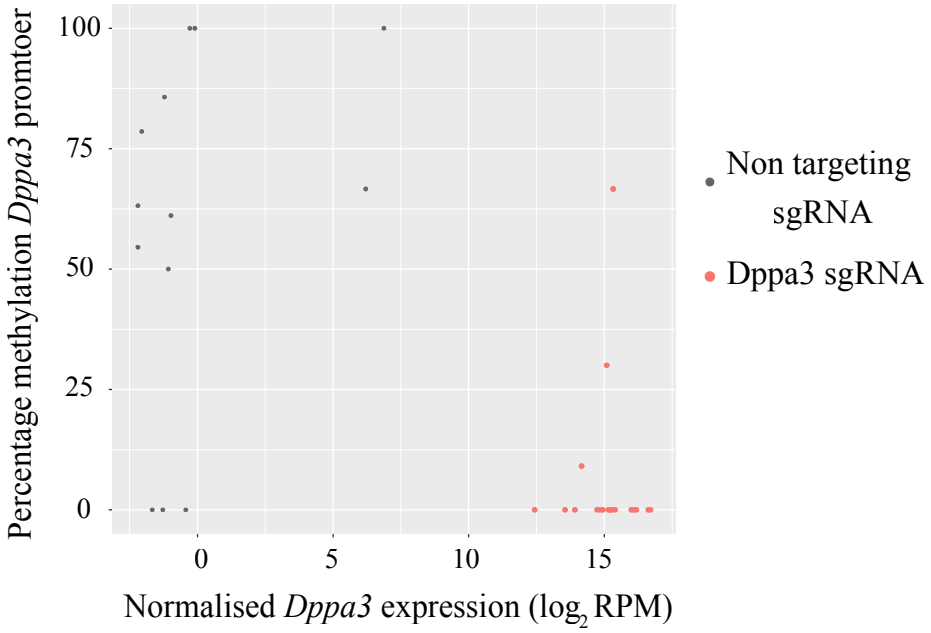



**Figure 4.41– Correlation between *Dppa3* transcriptional activation by CRISPRa and *Dppa3* promoter DNA methylation**

Scatterplot between normalised expression of *Dppa3* in $\log_2$ reads per million (RPM) (x axis) and percentage DNA methylation at the *Dppa3* promoter (y axis) in SAM22 mESCs transduced with Dppa3 sgRNA 379 (light red) or with non-targeting sgRNA control 462 (grey) (see Appendix C), analysed by scNMT-seq. Each dot represents a single cell. Percentage DNA methylation in the *Dppa3* promoter was calculated as the percent of methylated CpGs over total CpGs in the region 1.5 kilobases upstream and 0.5 kilobases downstream of the *Dppa3* TSS, without a threshold for minimum number of observations.

## 4.3 Conclusions and discussion

In this chapter, I describe the steps I took to select relevant candidate genes to screen for regulators of ZGA, by identifying epigenetic and transcriptional factors whose proteins are present in MII oocytes (Figures 4.1 and 4.2). Subsequently, a lentiviral library containing sgRNAs to target the promoters of these screen candidates was generated and transduced at low MOI into SAM22 mESCs (Figures 4.5-4.9). After scRNA-seq of the pool of transduced cells and generation of amplicon sgRNA libraries (Figure 4.10), I assigned sgRNA expression to each individual cell (Figure 4.13) and assessed the effectiveness of target gene activation by the CRISPRa SAM system (Figures 4.18-4.41). Consistent with the immunofluorescence and scRNA-seq results of chapter 3 (Figures 3.7, 3.21, 3.22 and 3.31), I noted a wide range in the

CRISPRa efficiency of target gene activation at the single-cell level (Figure 4.18); however, I found this could not be explained by expression of the sgRNA (Figure 4.22). Furthermore, I categorised the 460 transduced targeting sgRNAs as effective and ineffective as well as the 230 target genes as activated and not-activated (Figure 4.20). While I could partially underpin the cause of sgRNA effectiveness to the targeting position within the gene promoter (Figure 4.23), I could not find a single factor that, on its own, explained why some target genes could not be activated (Figures 4.26-4.34). In this section, I discuss the relevance of these results in relation to published literature, speculate about different possibilities to interpret these results and propose further experiments and/or analysis that could help towards the understanding of the molecular mechanisms governing CRISPRa.

In the first description of CROP-seq as a screening tool using pooled CRISPR KO libraries (Datlinger et al. 2017), the authors applied computational downsampling analysis to show that only 12-13 cells per sgRNA can suffice to detect the expected transcriptional signatures upon deletion of T-cell receptor signalling regulators. In chapter 3 of this dissertation, I performed *a priori* power calculations to estimate that at least 367 cells were required per sgRNA to confidently detect a ZGA-like transcriptional response, with the assumption that a positive hit would have similar effects to MERVL and *Zscan4* CRISPRa (Figure 3.24). While these design choices are highly dependent on the biological context under investigation, it is crucial to determine these parameters using prior knowledge or preliminary data. For instance, these preliminary tests allowed me to calculate the total number of cells to be sequenced to cover the 230 perturbations with enough statistical power (Figure 4.4). In fact, after merging single-cells across the three transduction replicates, which are highly reproducible (Figure 4.17), the dataset obtained for analysis consisted, on average, of 437 cells per sgRNA (Figure 4.15), surpassing the number of 367 cells per sgRNA estimated in chapter 3 to reach statistical power. However, the number of cells captured expressing each sgRNA ranged from 9 to 3,300, following the distribution of sgRNA representation in the plasmid library before transduction (Figure 4.16, Appendix C) and, therefore, for some sgRNAs, less than 367 cells were captured even after merging replicates (Figure 4.15, Appendix C). This raises the question of whether, for these sgRNAs, there was going to be enough statistical power to detect a ZGA signature in case they triggered one. Consequently, when identifying potential hits (see chapter 5), one needs to pay

particular attention on cell numbers and be aware that false negatives might arise as a consequence of lack of statistical power. Nevertheless, it is important to remember that the 367 cells per sgRNA were calculated on the basis of MERVL and *Zscan4* CRISPRa triggering a ZGA-like signature in ~10% of cells. However, and as I already discussed in chapter 3, it is possible to speculate that a ZGA regulator independent on the MERVL/*Zscan4*-regulated transcriptional pathway could be stronger in the induction of a ZGA-like signature in mESCs and, therefore, the triggered transcriptional signature would be detected with less than 367 cells.

On the technical aspect, 78.6% of the cells that passed quality control were also detected in the amplicon libraries (see Materials and Methods). Importantly, the 96% capture efficiency rate reported in the pilot test presented in chapter 3 (Figure 3.34) was calculated on the basis of amplified 10X Genomics cell barcodes in the amplicon libraires that could be associated to a sgRNA, without considering captured cell barcodes in the whole-transcriptome data. Consequently, the assignment rate after merging the amplicon data with whole-transcriptome libraries and quality controls is lower in the screen presented in this chapter. For the remaining 21.4% of cells whose cell barcode was not captured in the amplicon sgRNA libraries, it could be due to one or several of the following reasons: 1) the sgRNA was expressed but it was not captured (drop-out) with in the oligo(d)T-coated beads; 2) the sgRNA was expressed and captured with the oligo(d)T-coated beads but was not sufficiently amplified in the amplicon sgRNA libraries; or 3) more unlikely, the cells did not express a sgRNA despite surviving puromycin selection.

From the cells that passed quality control and were detected in the amplicon libraries, 99.54% could be assigned to one or multiple sgRNAs (Figure 4.13). In a small percentage of cell barcodes captured in the amplicon libraries (0.46%), a sgRNA sequence from those included in the pooled lentiviral library could not be confidently assigned (Figure 4.13). This is due PCR and sequencing errors in the sgRNA amplicon libraries (see Materials and Methods). More than 80% of cells were assigned to expression of a unique sgRNA (Figure 4.13). This percentage is lower than what was achieved in the pilot test (Figure 3.34). However, this is not surprising as, in the pilot test, sgRNAs were individually transduced and cells from different transductions only pooled together at the time of sequencing. This means that if a cell was

assigned to multiple sgRNAs, it was due to either the formation of cell doublets at the time of droplet generation in the 10X Genomics chromium controller or due to the formation of transcript chimeras during amplicon sgRNA library preparation, but not due to multiple infection events. On the contrary, detection of multiple sgRNAs in a cell after transduction of a pooled sgRNA library could be the result of cell doublets, transcript chimera formation or multiple infection events. Considering that the probability of multiple infection events at a MOI of 0.1 is minimal (Figure 4.6), it is likely that most of the cells assigned to two or more sgRNAs are due to noise introduced by chimeric transcripts in the amplicon libraries or due to doublets or "multiplets" formed during scRNA-seq droplet formation. Importantly, the doublet rate in 10X Genomics scRNA-seq libraries increases with the number of cells loaded (https://support.10xgenomics.com/single-cell-gene-expression), which was considerably higher in the screen than in the pilot test. Several computational and experimental methods have been developed in recent years to discriminate doublets in scRNA-seq libraries (Stoeckius et al. 2018; Kang et al. 2018; Gehring et al. 2018; C. Guo et al. 2018; Rosenberg et al. 2018; McGinnis et al. 2019). The results presented in this chapter indicate that the introduction of unique genomic barcodes can be used as a robust tool for doublet detection, which is consistent with a previous study (C. Guo et al. 2019).

To assess target gene activation, a method was developed to detect statistically-confident activated cells for the target gene compared to non-targeting sgRNA control cells (Figure 4.18, Appendix C, see Materials and Methods). Since this classification of activated and not-activated cells might be influenced by technical scRNA-seq drop-outs, I established a somewhat relaxed cut-off of 10 activated cells to call a sgRNA effective and classify the candidate genes as activated or not-activated in the dataset (Figure 4.20). The sgRNAs included in the lentiviral library were previously designed to increase on-target activity and decrease off-target effects (Konermann et al. 2015; Joung et al. 2017), however, I found a wide range of CRISPRa efficiencies after applying this cut-off (Figure 4.20). The most striking observation was that, within activated genes and a window of 180 bp upstream of the gene TSS, ineffective sgRNAs mapped closer to the TSS than effective sgRNAs (Figure 4.23). Most effective sgRNAs mapped to a window of 50-140 bp upstream of the TSS (Figure 4.23), further narrowing down the window for optimised efficiency in CRISPRa applications in comparison

to the optimal 0-200 bp window that others have previously described (Konermann et al. 2015; Horlbeck et al. 2016). These results are however in agreement with a recently optimised CRISPRa library containing sgRNAs targeting the region 75-150bp upstream of the gene TSS (Sanson et al. 2018). Validation experiments, as well as larger CRISPRa sgRNA libraries that can be assessed using a transcriptomic read-out, are needed to further confirm this targeting preference. Moreover, previous studies have shown that annotation of TSSs with the FANTOM database (Forrest et al. 2014), which identifies TSSs using Cap Analysis of Gene Expression (CAGE), is more accurate for sgRNA design in both CRISPRi and CRISPRa application than RefSeq or Ensembl annotations (Radzisheuskaya et al. 2016; Horlbeck et al. 2016; Sanson et al. 2018). Given that the Joung et al. 2017 library that I used to generate my custom sgRNA library used TSS RefSeq annotation, it would be important to analyse the sgRNA distance to the TSS using a mESC-specific FANTOM database to see if the differences observed for efficient and ineffective sgRNAs hold true.

Even though sgRNA GC content has been shown to be critical for CRISPR applications (T. Wang et al. 2014; Radzisheuskaya et al. 2016; Chari et al. 2015; Doench et al. 2016; H. Xu et al. 2015; X. Xu et al. 2017; Graf et al. 2019), it was expected to find no differences between effective and ineffective sgRNAs in the dataset (Figure 4.24), since they were all pre-selected to have an optimal GC content (Konermann et al. 2015; Joung et al. 2017). However, I observed some subtle differences in base composition in positions 8, 15 and 17 (Figures 4.25 and 4.26). These differences are in agreement with previous studies showing that sgRNA PAM-proximal bases influence on-target (Konermann et al. 2015; Joung et al. 2017) and off-target activity (Hsu et al. 2013; B. X. H. Fu et al. 2014; T. Wang et al. 2014; Chari et al. 2015; Graf et al. 2019), with one study specifically showing that position 17 determines the precision of DNA editing in CRISPR KO applications (Chakrabarti et al. 2019). It is therefore possible to speculate that the bases in positions 8, 15 and 17 are important for on-target dCas9-sgRNA binding in CRISPRa applications. However, similar to the observation made for the sgRNA targeting position upstream of the gene TSS, these observations are limited by the relatively small sgRNA library size used in this study and, therefore, need to be corroborated with larger libraries and independent validations. Importantly, given also the variety of CRISPRa methods (Chavez et al. 2016) (Figure 1.16), it will be crucial to test whether these rules are universal for

CRISPRa or method-dependent and whether they also apply to other CRISPR-editing applications (Figure 1.15). Furthermore, multiple algorithms have been developed in recent years taking into account sgRNA base content, targeting position and other factors to aim to predict CRISPR efficiency in different applications (H. Xu et al. 2015; Horlbeck et al. 2016; Doench et al. 2016; Radzisheuskaya et al. 2016; Haeussler et al. 2016; Labuhn et al. 2017), although very few applicable to CRISPRa (H. Xu et al. 2015; Horlbeck et al. 2016) and none of them based on scRNA-seq read-outs. Nonetheless, it would be interesting to see whether these algorithms can predict the observed efficiency in my CRISPRa dataset.

Having underpinned the cause of sgRNA effectiveness to the targeting position and, possibly, to base content in certain nucleotides, I next asked whether there were also features of the targeted genes themselves that made them more or less susceptible to CRISPRa. The first evidence that indicated differences in gene susceptibility to CRISPRa was that, whereas the differences in the sgRNA targeting position, and partially base content, were clear between effective and ineffective sgRNAs for genes that were activated with at least one sgRNA, those differences were not as clear for sgRNAs targeting genes that were not activated by either sgRNA (Figures 4.23-4.26). The basal level of gene expression was similar between activated and not-activated genes when analysing bulk RNA-sequencing data from untransduced mESCs (Figure 4.27), but consistently higher in activated genes when analysing scRNA-seq data from different library preparation methods (Figure 4.28). Whereas this contradicts previous observations that low levels of basal gene expression favour CRISPRa (Gilbert et al. 2014; Konermann et al. 2015), it is important to notice that most of the tested candidate genes are expressed at relatively high basal levels in mESCs (Figure 4.3). This is consistent with the high basal levels for ATAC-seq, DNase-seq, H3K4me3 and H3K27ac signals and low basal levels for H3K27me3, H3K9me3 and DNA methylation signals at these gene's promoters, as well as high basal levels of H3K36me3 in gene bodies, regardless of the activation state of the gene achieved by CRISPRa (Figures 4.30-4.34). These observations, together with the lack of differences in expression variability at the single-cell level (Figure 4.29), suggest that the higher basal expression detected for activated genes is probably a technical artifact of scRNA-seq libraries that makes those genes more easily detectable, consistent with drop-outs being more common for lower expressed genes (Qiu et al. 2018). If this was true, it raises the

possibility that genes that I classified as "not-activated" were actually activated but the level of activation was not detectable or did not pass the threshold of 10 activated cells.

Several epigenetic features have been shown to influence CRISPR efficiency, although there has not been any study so far addressing epigenetic influence in CRISPRa applications. An open chromatin state and low nucleosome occupancy generally favours both Cas9 and dCas9 binding (Kuscu et al. 2014; Xuebing Wu et al. 2014; Chari et al. 2015; Singh et al. 2015; Knight et al. 2015; Horlbeck, Witkowsky, et al. 2016; Isaac et al. 2016; Radzisheuskaya et al. 2016; Jensen et al. 2017; Uusi-Mäkelä et al. 2018; Yarrington et al. 2018). H3K27ac, H3K9ac, H3K4me2 and H3K79me2 have also been shown to favour CRISPR editing (Radzisheuskaya et al. 2016; Chari et al. 2015) and DNA CpG methylation was paradoxically reported to affect CRISPR mutagenesis in some studies but not in others (Xuebing Wu et al. 2014; Hsu et al. 2013; Perez-Pinera et al. 2013). In the categories that I established to study CRISPRa efficiency in my dataset, I could not find any epigenetic feature that individually explains the differences observed between activated and not-activated genes (Figures 4.30-4.34). This does not necessarily contradict previous studies, as none of them specifically analysed CRISPRa sgRNA libraries and also, it might be context- and/or tissue/cell type-dependent. It is possible that a combination of several of these features, but none of them individually, can predict CRISPRa efficiency. Therefore, in the future, it will become important to systematically analyse larger CRISPRa libraries in relation to the basal epigenetic state of the targeted genes in different cell types, and generate algorithms that can predict which feature or combination of features determine CRISPRa efficiency.

As mentioned earlier, an important consideration when interpreting these results is that I am setting a threshold for each sgRNA to consider it effective (10 activated cells) and therefore categorise the target gene as activated. Nevertheless, one could speculate that effective activation of less than 10 cells is also sufficient to trigger a transcriptional response downstream of the upregulation of the targeted gene. Consequently, it is possible that sgRNAs that were categorise as ineffective were actually effective to trigger an activation of the endogenous target sufficient to detect a downstream transcriptional response. If this was the case, the signal-to-noise ratio in the epigenetic features analysed for the groups of activated and not-activated genes might be too low to detect statistically significant differences. Another possibility is that

the dynamics of target gene activation is gene dependent, as shown in Yanxia Liu et al. 2018, and, therefore, one could imagine that, before the cells were analysed by scRNA-seq 10 days after transduction of the lentiviral library, some target genes had a peak of activation during that time course that led to downstream transcriptional effects, but target gene activation was not retained until day 10, therefore uncoupling target gene activation and downstream response. Consequently, when calling screen hits (see chapter 5), it will be important to consider that a sgRNA could have induced a ZGA-like response in mESCs even if activation of the target gene is not strikingly observed.

At the single-cell level, I already discovered in chapter 3 that there was variability in target gene activation (as analysed for MERVL and *Zscan4*) between single-cells transduced with the same effective sgRNA (Figures 3.7, 3.21, 3.22 and 3.31). From those immunofluorescences and scRNA-seq experiments, I could not tell whether the cells were expressing the sgRNA to conclude whether there was real heterogeneity. In this chapter, I could confirm these observations for multiple sgRNAs and target genes (Figures 4.18, 4.19 and Appendix C). Since Cas9/dCas9 and sgRNA expression levels have been shown to be critical for effective CRISPR in multiple studies (Hsu et al. 2013; Doench et al. 2014.; Labuhn et al. 2017; Yuen et al. 2017), I used the information derived from the sgRNA amplicon libraries to confirm that sgRNAs were not differentially expressed in activated and not-activated cells (Figure 4.22). Unfortunately, due to technical reasons, I could not assess whether this was also the case for dCas9/VP64 and/or MS2-p65-HSF1 expression (Figure 4.21), although it is unlikely that these two fusion proteins were expressed differentially between cells given the clonal nature of the SAM22 cell line (Figure 3.3). As I discussed in chapter 3, the other alternative that could explain variability between single-cells in target gene activation by CRISPRa would be the presence of epigenetic heterogeneity at the target gene promoter. mESCs present transcriptional and epigenetic heterogeneity (Rulands et al. 2018; Luo et al. 2018; Clark et al. 2018; reviewed in Tanaka 2009) and, therefore, it was possible to think that this heterogeneity determined the ability of single-cells to activate the target gene even when all the CRISPRa machinery was expressed. To test this hypothesis, I used scNMT-seq on *Dppa3* CRISPR-activated cells, aiming to find a correlation between *Dppa3* promoter accessibility and/or DNA methylation and the *Dppa3* level of expression triggered in each cell by CRISPRa. However,

due to the sparsity of the data, very few cells had enough reads at the *Dppa3* promoter in the single-cell bisulfite libraries to be able to do confident analysis, making this experiment less than ideal to address this question. Furthermore, even if a correlation between target gene activation and chromatin accessibility and/or DNA methylation at the target gene promoter had been observed, it would have been difficult to assess the causality of the events, i.e, does a certain epigenetic state at the target gene promoter determine effective target gene activation by CRISPRa, or is the epigenetic state of the promoter observed after CRISPRa a consequence of the targeting? scNMT-seq could not have answered this question since it requires lysis of the cells for analysis (Clark et al. 2018). A live cell imaging system where the expression of the target gene as well as the promoter epigenetic state can be reported before and after induction of CRISPRa is the only way to address such questions and, to the best of my knowledge, there is currently no technology to perform an experiment of this kind. Alternatively, and similarly to what was proposed above, it is possible to think about a cycling effect in target gene activation, where some cells silence the gene after having induced it with CRISPRa. This dynamic expression could also be dependent on cell cycle and, to test this hypothesis, future analysis should assess the cell cycle state of activated versus not-activated cells.

To conclude, in this chapter, I generated a dataset of single-cells with individual CRISPRa perturbations and found that CRISPRa efficiency was sgRNA- and gene-dependent. While I analysed multiple features to try to explain these differences in efficiency, I only found a significant effect in the targeting position of the sgRNA. Larger studies using more sgRNAs as well as different biological contexts are needed to systematically identify the targeting rules of CRISPRa. For the next chapter of this dissertation, where I identify screen hits for ZGA regulators, there are two important considerations that derive from the conclusions presented here: 1) detection of target gene activation might be uncoupled to downstream transcriptional responses and, therefore, it is possible to identify screen hits where no target gene activation is observed, and 2) due to this wide range in efficiency, it is possible to obtain only one sgRNA as a hit out of the two included for the target gene, lowering the confidence in the identification of hits. For both of these reasons, it becomes crucial to validate screen hits individually and with alternative approaches.

# Chapter 5 Identification and molecular characterisation of regulators of ZGA

## 5.1 Background and summary

In the previous two chapters, I set up a CRISPRa method in mESCs with scRNA-seq read-out (chapter 3), which I subsequently applied to screen 230 maternal epigenetic and transcriptional factors (chapter 4). In this final chapter of results, I collaborated with Danila Bredikhin and Oliver Stegle (from EMBL, Heidelberg, see Table of Acknowledgement of Assistance) to implement a computational method to identify screen candidates that activated a ZGA-like transcriptional signature in the CRISPRa scRNA-seq dataset that I generated. This method used an integrative dimensionality reduction approach of the expression of coding genes as well as repetitive elements in each single cell based on multi-omics factor analysis models (MOFA) (Argelaguet et al. 2018). MOFA was used to infer a set of "factors" that captured expression variability within the dataset, one of these factors being a ZGA-like factor which was subsequently used to identify 44 screen hits.

Amongst the identified hits for ZGA regulation were the DNA binding protein *Dppa2*, the chromatin remodeller *Smarca5* and the POZ-, AT hook-, and zinc finger protein-1 (*Patz1*). I independently validated these three hits, together with the negative control Calcium-Regulated Heat Stable Protein-1 (*Carhsp1*), by individual CRISPRa experiments as well as cDNA-eGFP overexpression, both followed by bulk RNA-sequencing. These validation experiments further confirmed that, upon transcriptional upregulation of these factors by either overexpression method, mESCs adopted an early embryonic-like state that resembled the transcriptional profile of mid-to-late two-cell embryos.

Finally, through a series of KO and overexpression experiments, I disentangled part of the ZGA regulation network by showing that *Smarca5* regulates the expression of ZGA genes via *Dppa2*. In turn, *Dppa2* mediates its effects via the previously described transcription factor *Dux* (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Y.-L. Yan et al. 2019).

## 5.2 Results

### 5.2.1 CRISPRa of selected candidates induced a ZGA-like transcriptional response in mESCs

Having generated a high-quality scRNA-seq dataset of perturbed cells by CRISPRa of maternal epigenetic and transcriptional factors, I first addressed whether this targeting strategy had worked for inducing a ZGA-like signature in SAM22 mESCs. To this end, I used the PCA performed in chapter 4 on highly-variable genes amongst the 203,894 single-cells expressing a unique sgRNA (Figure 4.17) to understand the major sources of variation in the dataset. Gene ontology analysis of the top 50 gene loadings for the first component (PC1), using all highly variable genes in the dataset as a background (see Materials and Methods), suggested that this component captured intrinsic variance in cell shape and cell contacts (Appendices E and F), likely reflective of different pluripotent and differentiation states of the cells upon CRISPRa of different screen candidates. Enriched gene ontology terms for the top gene loadings for PC1, although with relatively high corrected p-values (or FDR), included "actomyosin structure organization" (FDR = 0.0022), "actin filament-based process" (FDR = 0.34), "sarcomere organization" (FDR = 0.29) or "cell adhesion" (FDR = 0.22), with markers such as the calponin variants *Cnn1* or *Cnn2*, which are involved in actin-myosin regulation (Boraas et al. 2018), the keratin proteins *Krt8* or *Krt19,* which are key fibrous structural proteins involved in mESC differentiation (Maurer et al. 20108), or actin gamma cytoplasmic-1 (*Actg1*), also involved in cytoskeletal reorganization during cell adhesion and differentiation (reviewed in Ambriz et al. 2018) (Appendices E and F). In contrast, gene ontology analysis of the top 50 gene loadings for the second component (PC2) did not reveal any enriched functional processes other than telomeric regulation (FDR = 0.28) with *Zscan4* genes as markers (Appendices E and F).

Since ZGA genes are largely uncharacterised and, consequently, they are not generally associated with gene ontology functions (Eckersley-Maslin et al. 2016), I decided to analyse whether the top gene loadings for PC2 were enriched in genes expressed during the major wave of ZGA. To this end, I constructed a list of ZGA genes by merging different relevant published datasets, including the MERVL/*Zscan4*-driven network described in Eckersley-Maslin et al. 2016, *Dux*-regulated genes as described in Hendrickson et al. 2017, and other transcripts expressed in two-cell mouse embryos, as recently reported in Y Li et al. 2018 (Appendix I). Excitingly, comparison of the top gene loadings for PC1 and PC2 to this ZGA gene list revealed that PC2, but not PC1, captured expression variation of a ZGA-like transcriptional signature (Figure 5.1, Appendix E, Appendix I).

To further verify this, I next looked at the expression of ZGA markers, such as *Zscan4c*, *Zscan4d*, *Gm8300* or *Tmem92*, in cells plotted along PC1 and PC2 axis and confirmed that the cells contributing to PC2 variance had higher expression of these markers (Figure 5.2). Furthermore, while the top 50 gene loadings for PC1 are highly expressed at the blastocyst stage, reflecting the transcriptome of serum-grown mESCs, the top 50 gene loadings for PC2 showed a peak of expression in mid-to-late two-cell embryos, when the major wave of ZGA occurs (Figure 5.3), confirming that PC2 captured a ZGA-like signature in the dataset.

Importantly, despite the variance captured in PC1, which might be indicative of cell differentiation (Appendices E and F), the pluripotecy marker *Pou5f1* (also known as *Oct4*) showed high and uniform expression across the majority of cells in the dataset (Figure 5.4). Consistently, it was not detected as a highly variable gene (Appendix E). This is also in agreement with the high corrected p-values of enriched gene ontology terms for the top gene loadings for PC1 (Appendix F) and with previous reports showing that 2C-like cells lack OCT4 protein but the mRNA is not downregulated (Macfarlan et al. 2012; Ishiuchi et al. 2015; Eckersley-Maslin et al. 2016; Choi et al. 2017; Rodriguez-Terrones et al. 2018).

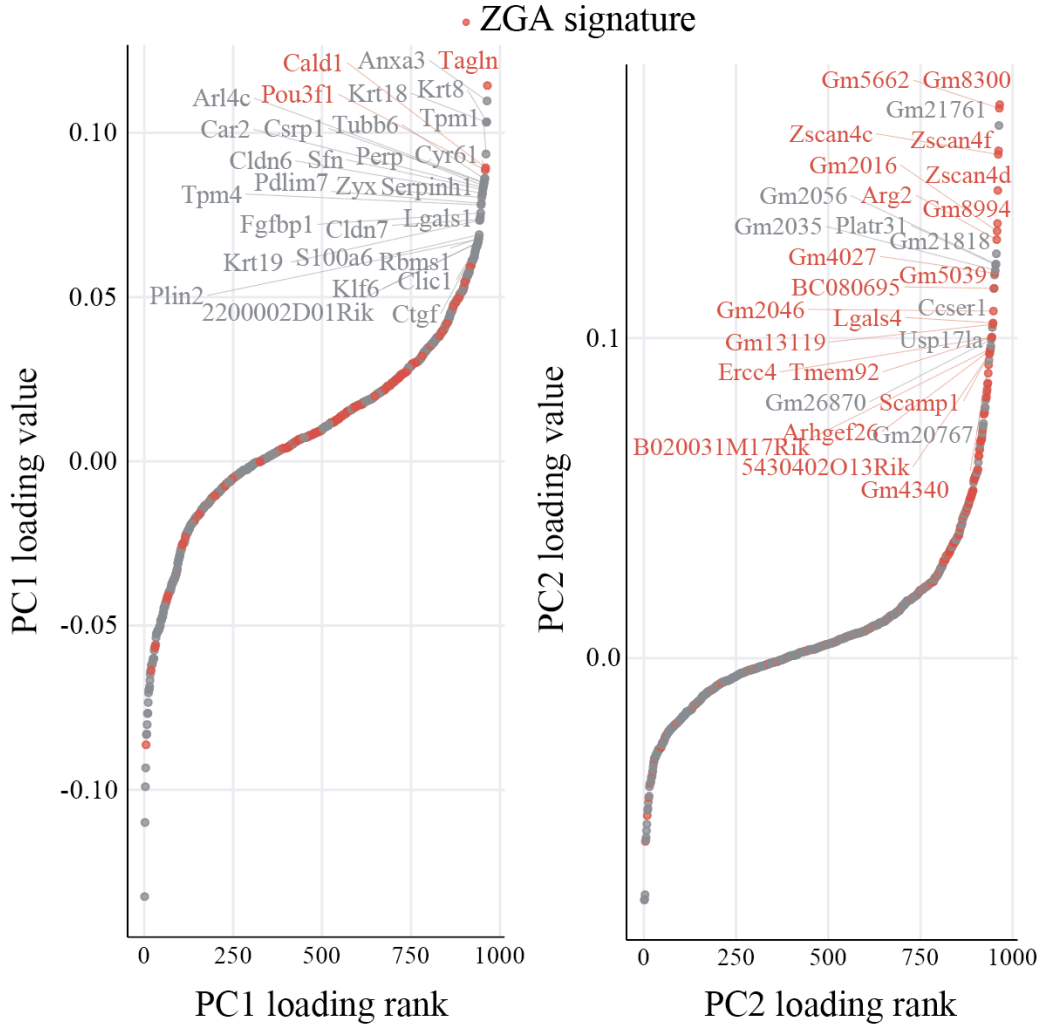



**Figure 5.1– Highly variable genes ranked by their loadings in PC1 and PC2**

965 highly variable genes ranked by their loading value in principal component 1 (PC1) (left) and principal component 2 (PC2) (right), highlighting in red genes previously known to be expressed during the major wave of ZGA or ZGA-like transcriptional responses, as described in Appendix I (Eckersley-Maslin et al. 2016; Hendrickson et al. 2017; Y. Li et al. 2018). The loading value for each gene in each PCA component is available in Appendix E.

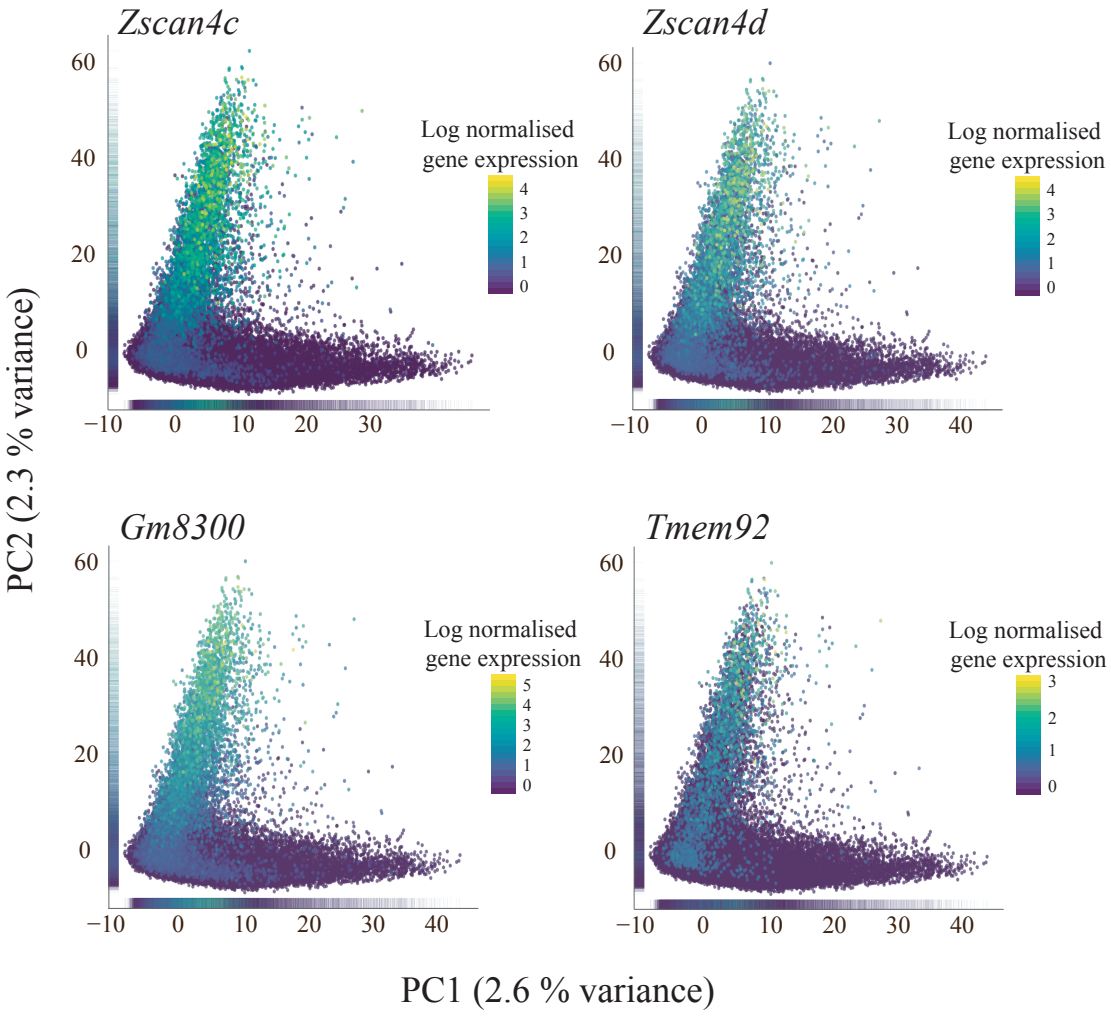



**Figure 5.2– Expression of ZGA markers along PC1 and PC2**

Visualisation of cells along principal component 1 (PC1, x axis) and principal component 2 (PC2, y axis) coloured by the logarithmic normalised expression of the ZGA markers *Zscan4c* (top left), *Zscan4d* (top right), *Gm8300* (bottom left) and *Tmem92* (bottom right). Marginal distributions of PC1 and PC2 values are displayed as rug plots along the respective axis. PC1 explains 2.6% of total variance in the whole dataset whereas PC2 explains 2.3% of variance.

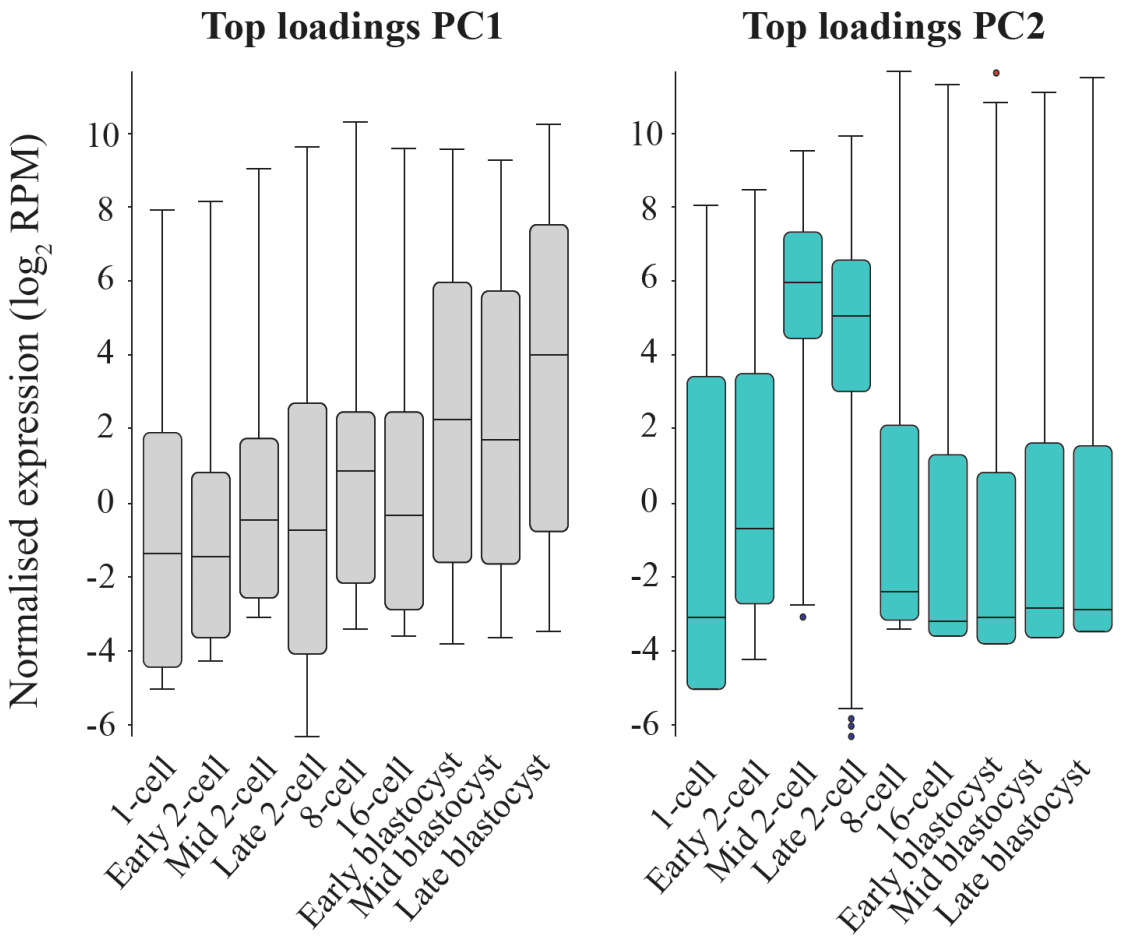



**Figure 5.3– Expression of top loadings for PC1 and PC2 during pre-implantation development**

Box-whisker plots showing normalised expression in $log_2$ reads per million (RPM) of the top 50 gene loadings for PC1 (grey - left) and the top 50 gene loadings of PC2 (light blue - right) during mouse pre-implantation development (data analysed from Deng et al. 2014). The loading value for each gene in each PCA component is available in Appendix E.

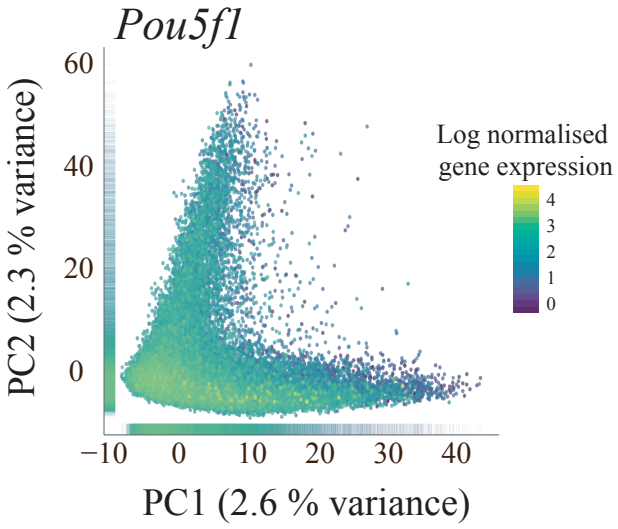



**Figure 5.4– Expression of *Pou5f1* along PC1 and PC2**

Visualisation of cells along principal component 1 (PC1, x axis) and principal component 2 (PC2, y axis) coloured by the normalised expression of the pluripotency gene *Pou5f1* (*Oct4*). Marginal distributions of PC1 and PC2 values are displayed as rug plots along the respective axis. PC1 explains 2.6% of total variance in the whole dataset whereas PC2 explains 2.3% of variance.

Next, I set to characterise the observed ZGA-like transcriptional signature in more detail. As introduced in chapter 1 (section 1.1.3.3), transposable or repeat elements are key drivers of gene expression during early embryonic development (reviewed in Rodriguez-Terrones & Torres-Padilla 2018). To account for their expression in the screen dataset and identify factors that induced a ZGA-like response both in the protein-coding and non-coding transcriptome, we trained a multi-omics factor analysis (MOFA) model (Argelaguet et al. 2018). This approach allowed treating the coding transcriptome and repeat elements as distinct "views" and identify "MOFA factors" that explain the transcriptional variability within the dataset (Figure 5.5, see Materials and Methods). Moreover, sgRNA-cell assignment information was inputted to the model (Figure 5.5, see Materials and Methods). We included eight relevant repeat families in the analysis whose expression could be detected in the dataset: LINE-1, ERVK, SINE B2, ERV1, MERVL, SINE B4, SINE Alu B1 and major satellites, compared to other repeat elements such as telomeres or ribosomal RNA (Figure 5.6 and see Materials and Methods).

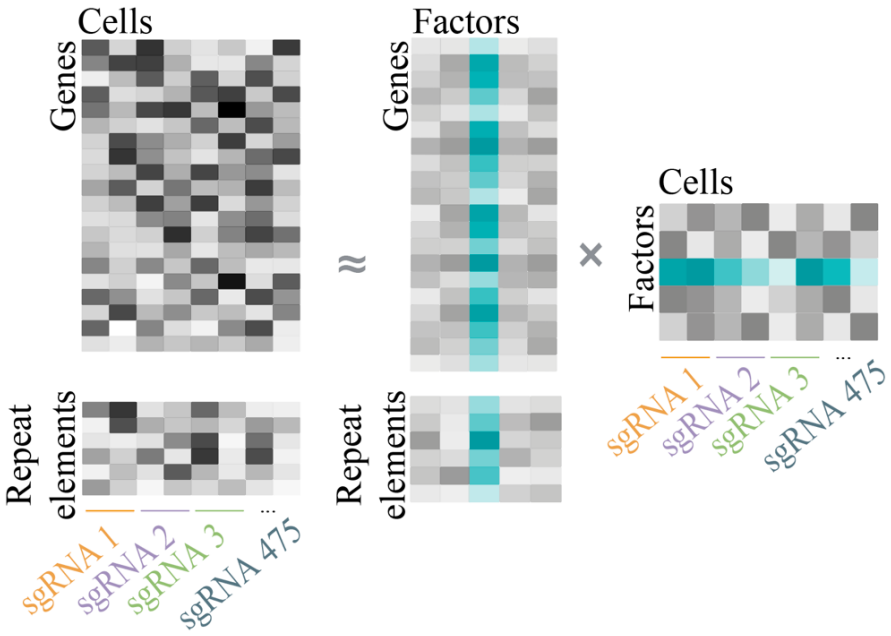



**Figure 5.5– Schematic of a MOFA model to identify screen hits**

Schematic of the joint analysis of protein-coding gene and repeat elements expression in the CRISPRa scRNA-seq dataset using multi-omics factor analysis (MOFA). Matrices of dimension features (genes or repeat elements) in cells grouped by sgRNA expression were treated as distinct views in the model and decomposed into the product of weights (or loadings) and factors. Factor 3 in the trained model, interpreted as a ZGA-like factor (see Figures 5.7-5.9), is highlighted in green.

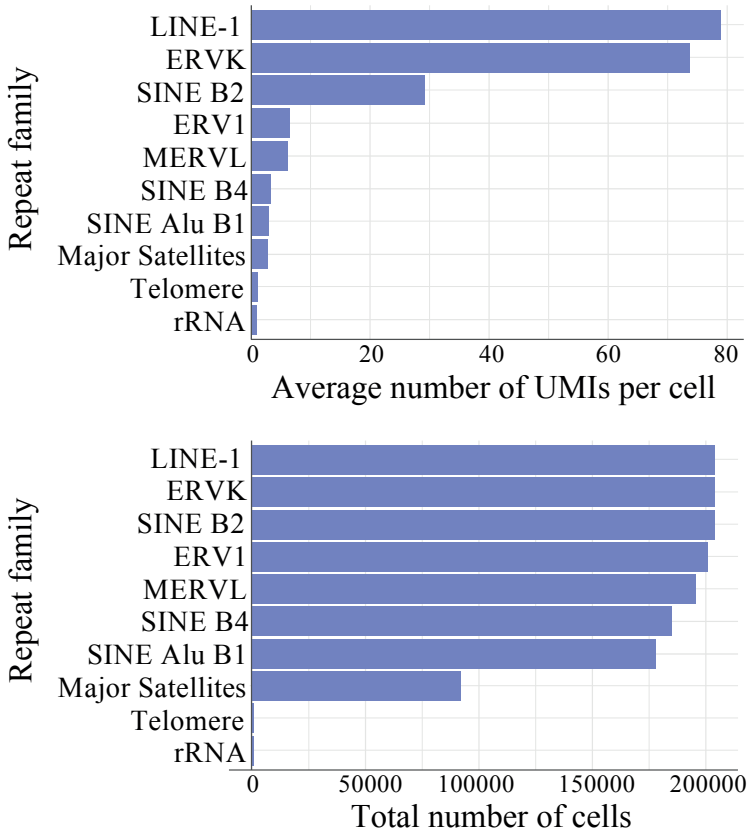



**Figure 5.6– Expression of repeat families in the CRISPRa scRNA-seq dataset**

**Top panel**: Average number of unique molecular identifiers (UMIs) per cell in the dataset for each of the ten repeat families analysed.

**Bottom panel**: Total number of cells in the dataset for which each repeat family is detected (UMI>0), out of a total of 203,894 cells analysed.

Excitingly, amongst the MOFA factors identified (Appendix G), the expression variance captured uniquely by factor 3 corresponded to a ZGA-like signature (Figure 5.5). This factor was identified as the ZGA-like factor for three reasons: 1) similar to PC2, the variable protein-coding genes that ranked top by their loadings in this factor, and not the top loadings for other factors, were enriched in ZGA or ZGA-like genes previously defined in published literature (Eckersley-Maslin et al. 2016; Hendrickson et al. 2017; Y. Li et al. 2018) (Appendix I) (Figure 5.7); 2) consistently, the top loading genes for this factor, and not for others, are highly expressed in mid-to-late two-cell embryos, at the time of the major wave of ZGA (Figure 5.8); and 3) the ZGA-associated MERVL (Macfarlan et al. 2012) and major satellite repeats (Casanova et al. 2013) were most prominently associated to factor 3 amongst the repeat classes analysed (Figure 5.9).

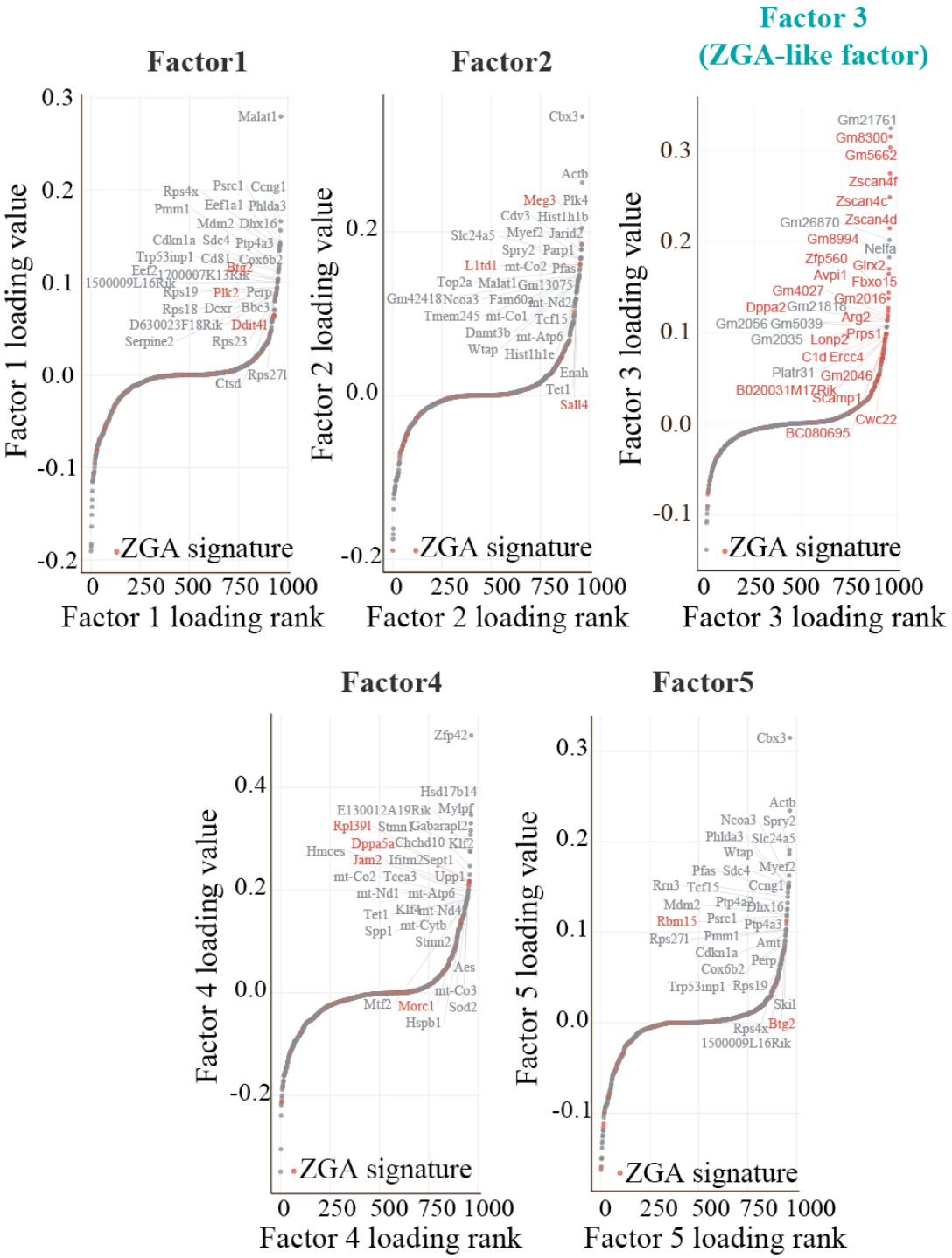



**Figure 5.7– Highly variable genes ranked by their loadings in MOFA factors 1-5**

965 highly variable protein-coding genes ranked by their loading value in MOFA factors 1-5, highlighting in red genes previously known to be expressed during the major wave of ZGA or ZGA-like transcriptional responses, as described in Appendix I (Eckersley-Maslin et al. 2016; Hendrickson et al. 2017; Y. Li et al. 2018). The enrichment of ZGA genes in factor 3 identifies it as the ZGA-like factor. The loading value for each gene in each MOFA factor is available in Appendix G.

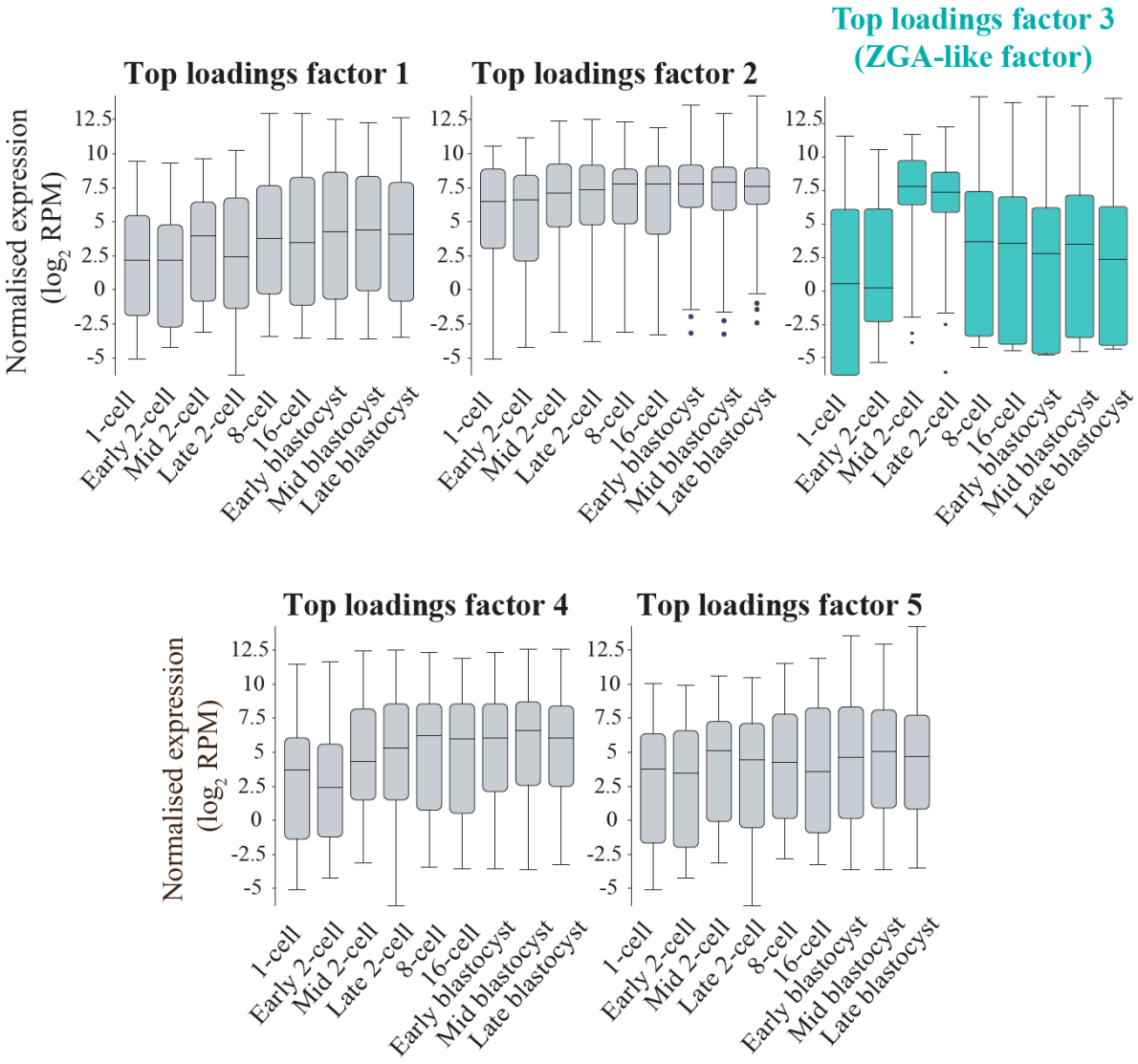



**Figure 5.8– Expression of top gene loadings for MOFA factors 1-5 during pre-implantation development**

Box-whisker plots showing normalised expression in $\log_2$ reads per million (RPM) of the top 50 protein-coding gene loadings for MOFA factors 1-5 during mouse pre-implantation development (data analysed from Deng et al. 2014). The high expression in mid-to-late two-cell embryos of the top loadings for MOFA factor 3 identifies it as the ZGA-like factor. The loading value for each gene in each MOFA factor is available in Appendix G.

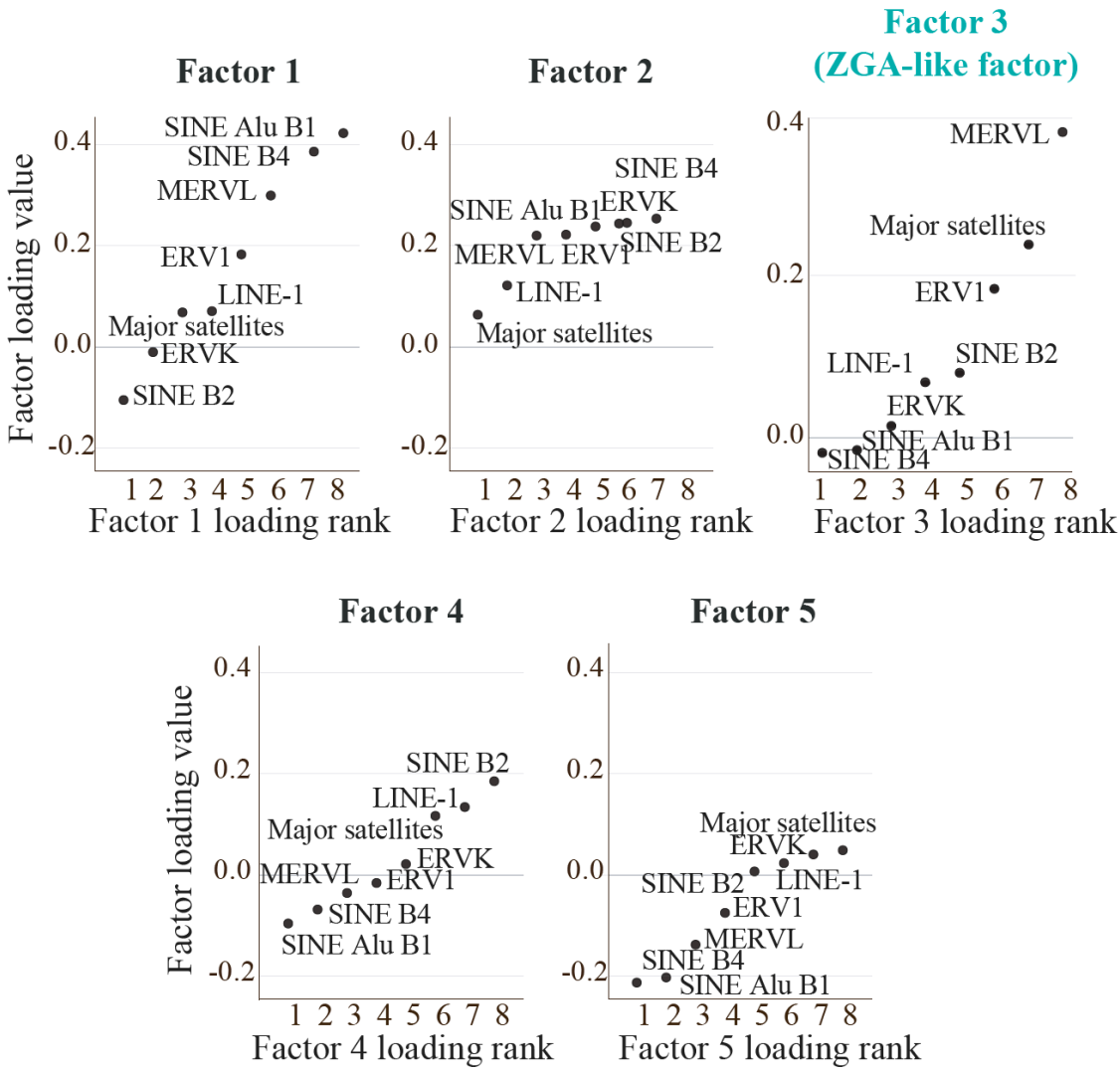



**Figure 5.9– Repeat element families ranked by their loadings in MOFA factors 1-5**

Repeat element families ranked by their loadings in MOFA factors 1-5. The high variance explained by MERVL and major satellite repeats in factor 3 identifies it as the ZGA-like factor.

Interestingly, unlike the top gene loadings for PC1 that showed a peak of expression in blastocyst stages (Figure 5.3), the top gene loadings for MOFA factors 1, 2, 4 and 5 showed high and uniform expression during all stages of pre-implantation development (Figure 5.8). This suggested that these factors captured biological or technical variability associated with mESC cultures rather than with specific transcriptional programmes in pre-implantation development. In fact, gene ontology analysis showed that factor 1 captured variability associated with protein metabolism and cell cycle; factor 2 is associated with epigenetic regulation, consistent with the epigenetic heterogeneity found in mESC cultures (Rulands et al. 2018; Luo et al. 2018; Clark et al. 2018; reviewed in Tanaka 2009); gene ontology analysis of factor 4 did not reveal any significant enrichment; and factor 5 seems associated with apoptotic and DNA damage responses, although the statistical significance of the enriched gene ontology terms was not very high, consistent with the removal of low-quality input during quality control of the dataset (Figure 4.12) (Appendices G and H, see Materials and Methods). Gene ontology analysis of the top loadings for factor 3 or ZGA-like factor did not reveal any enriched functional processes other than telomeric regulation with *Zscan4* genes as markers (Appendices G and H), same as what was observed for PC2 and consistent with the uncharacterised function of ZGA genes.

Altogether, this unsupervised analysis of the dataset indicated that CRISPR-activating maternal epigenetic and transcriptional factors in mESCs induced expression variation that mimics a ZGA-like transcriptional response, suggesting that a substantial fraction of the targeted candidates induced ZGA-like gene expression.

### 5.2.2 Identification of activators of a ZGA-like transcriptional signature

To reveal these screen hits whose CRISPRa induced ZGA-like expression, we grouped cells in the dataset by the expression of the targeting sgRNA and ranked each group based on its contribution to the overall transcriptional variance explained by the MOFA ZGA-like factor (Appendix C, see Materials and Methods). Since no major conclusions were obtained from the CRISPRa efficiency analysis performed in chapter 4 and, as discussed earlier, it is possible to speculate that target gene activation is uncoupled from downstream transcriptional responses, I decided to include all targeting sgRNAs in this ranking analysis, regardless of the target gene

activation state. This approach identified 46 sgRNAs targeting 44 unique candidate genes for which the MOFA factor explained a larger fraction of variance than for the 15 non-targeting sgRNA controls (Figure 5.10, Appendix C). At the top of the ranking were *Dppa2* and *Smarca5*, with both targeting sgRNAs as hits, and the transcription factor *Patz1* (Figure 5.10, Appendix C).

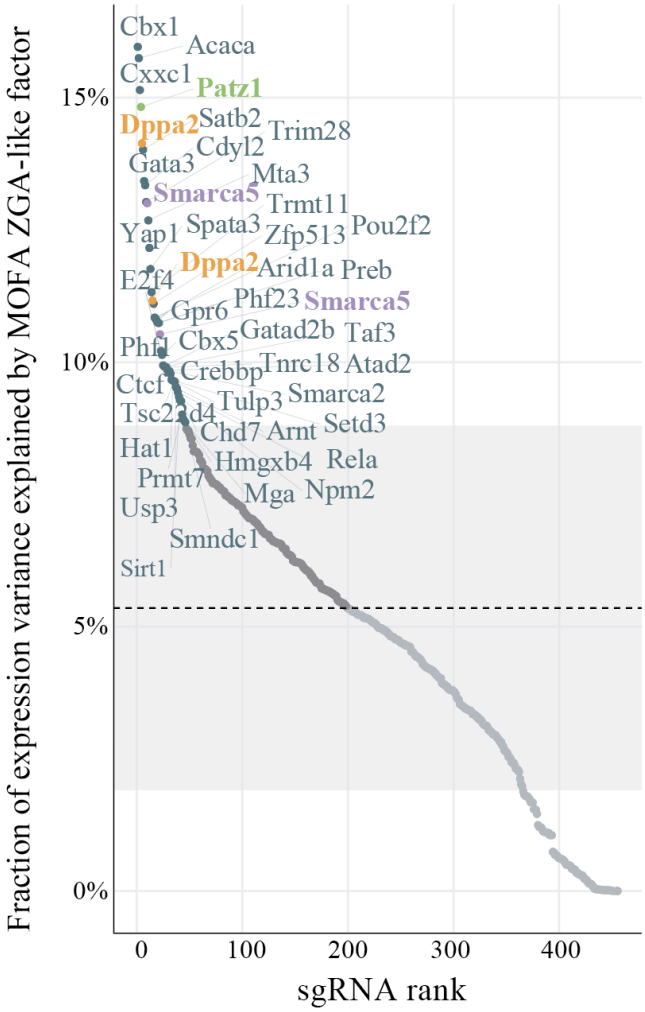



**Figure 5.10– Ranking of sgRNAs by their contribution to ZGA-like signature variance**

Ranking of the 460 targeting sgRNAs by the fraction of expression variance (average of protein-coding genes and repeat elements) explained by MOFA factor 3 or ZGA-like factor. The fraction of expression variance explained by MOFA factor 3 for the 15 non-targeting sgRNA controls is depicted in the background as mean (dashed line) and plus and minus one standard deviation (shaded area). The names of the 44 target genes whose sgRNAs were identified as hits because they exceed the variance explained by non-targeting sgRNA controls are labelled, with *Dppa2* (orange), *Smarca5* (purple) and *Patz1* (green) sgRNA hits highlighted. The fraction of variance and rank position for each sgRNA is available in Appendix C.

Importantly, SAM22 mESCs expressing these 46 sgRNAs not only showed an increased ZGA-like variance but also specifically induced the expression of ZGA genes associated with MOFA factor 3, while genes contributing to the variance of other MOFA factors remained largely unaltered between cells expressing the 46 sgRNA hits and cells expressing other targeting sgRNAs (Figure 5.11).

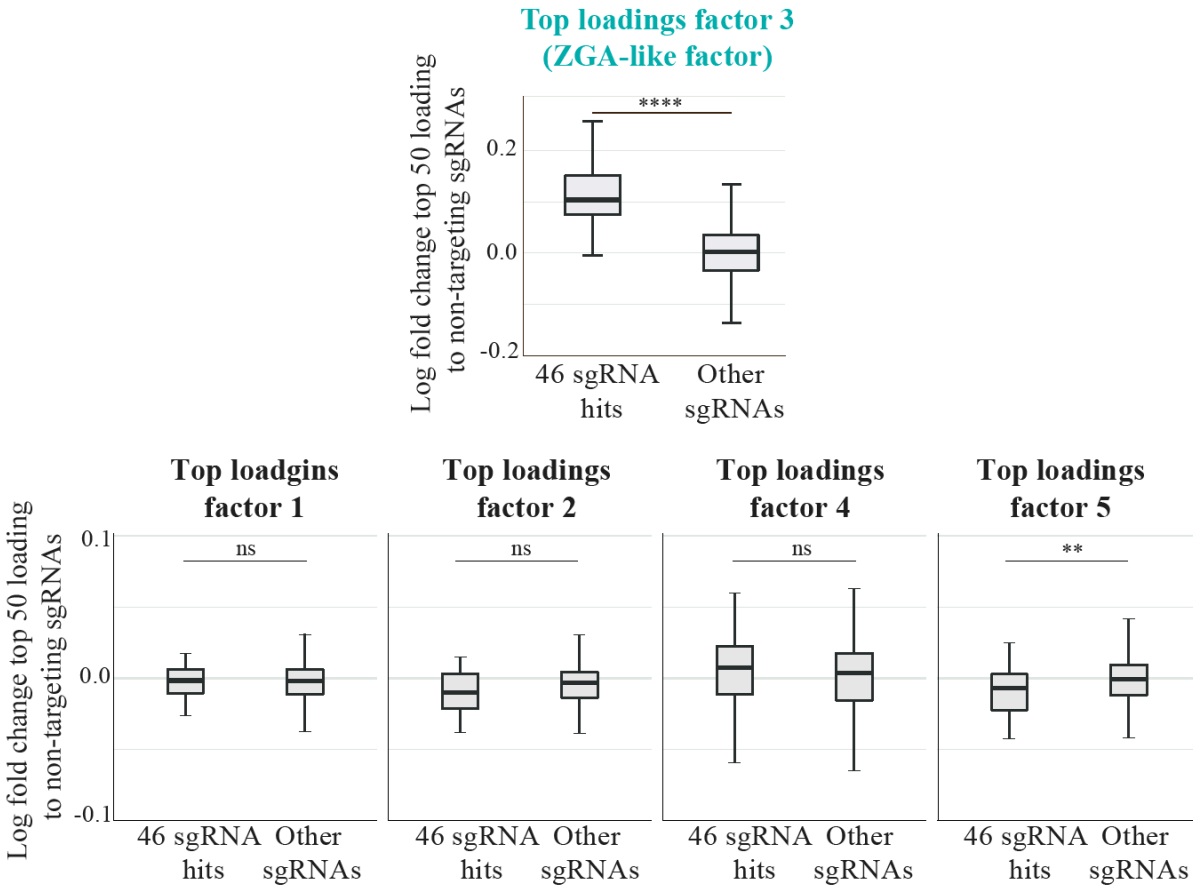



**Figure 5.11– Expression of top gene loadings for MOFA factors 1-5 in cells expressing sgRNA screen hits**

Box-whisker plots showing $\log_{10}$ fold change expression of the top 50 protein-coding gene loadings for MOFA factors 1-5 in SAM22 mESCs expressing the 46 sgRNA hits and SAM22 mESCs expressing other targeting sgRNAs, compared to SAM22 mESCs expressing non-targeting sgRNA controls. Expression was quantified in normalised counts. Statistically significant differences between the two groups were assessed by Mann-Whitney test: ****: p-value < 0.0001, **: p-value <0.01, ns: non- significant. The loading value for each gene in each MOFA factor is available in Appendix G. The list of the 46 sgRNAs identified as hits is available in Appendix C.

Similarly, MERVL and major satellite repeats were significantly upregulated in SAM22 mESCs expressing the 46 sgRNA hits (Figure 5.12), consistent with these repeat families ranking top in the MOFA ZGA-like factor (Figure 5.9). Interestingly, ERV1, SINE B2 and LINE-1 families, which followed MERVL and major satellites in the MOFA ZGA-like factor ranking (Figure 5.9), were also induced, although to a lesser extent, by the screen hits (Figure 5.12). Consistently, these transposon families have also been previously implicated in ZGA (W. Zhang et al. 2019; Vasseur et al. 1985; Percharde et al. 2018).

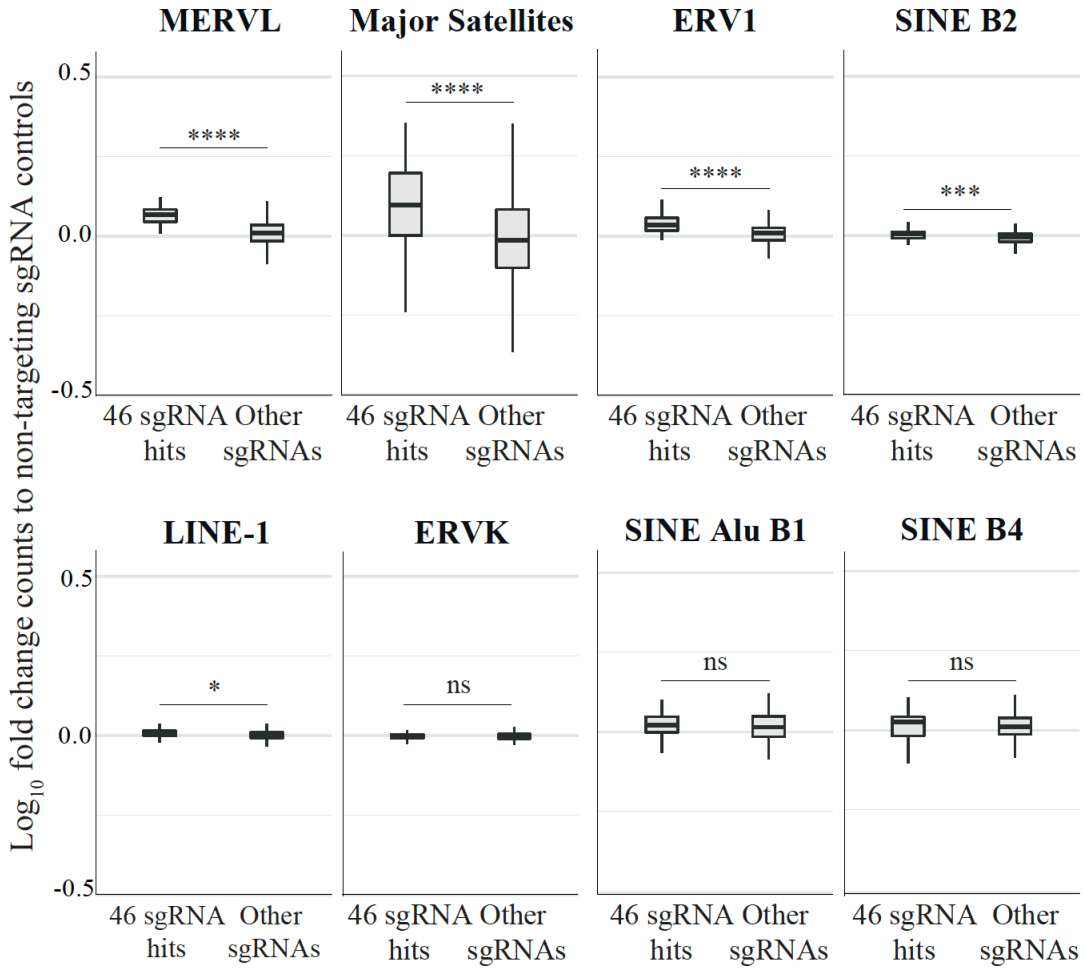



**Figure 5.12– Expression of repeat elements in cells expressing sgRNA screen hits**

Box-whisker plots showing $log_{10}$ fold change expression of different repeat families in SAM22 mESCs expressing the 46 sgRNA hits and SAM22 mESCs expressing other targeting sgRNAs, compared to SAM22 mESCs expressing non-targeting sgRNA controls. Expression was quantified in normalised counts. Statistically significant differences between the two groups were assessed by Mann-Whitney test: ****: p-value < 0.0001, ***: p-value <0.001, *: p-value <0.05, ns: non-significant. The list of the 46 sgRNAs identified as hits is available in Appendix C.

Next, I sought to understand the transcriptional changes triggered by each of these 46 sgRNAs. First, I analysed target gene activation. In chapter 4, I considered that a minimum of 10 cells out of all the cells that expressed a given sgRNA had to show target gene activation (FDR<10%, see Materials and Methods) to classify the sgRNA as effective. Strikingly, out of the 46 sgRNA hits, only 15 of them (32.6%) induced target gene activation in more than 10 cells (Figure 5.13).

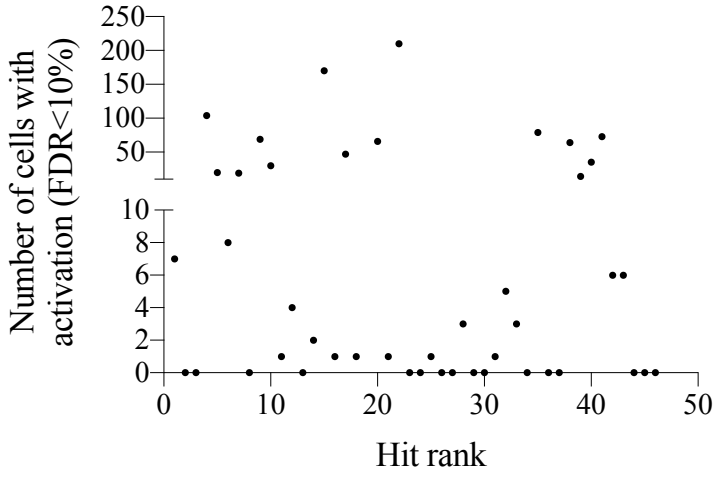



**Figure 5.13– Number of activated cells expressing sgRNA screen hits**

Number of activated cells (FDR<10%, as described in chapter 4 and Materials and Methods) ordered by MOFA sgRNA hit rank. Each dot represents the number of activated cells in each of the 46 sgRNA hits. The sgRNA hit rank is available in Appendix C.

This observation raised two alternative hypotheses: 1) the threshold of more than 10 activated cells considered in chapter 4 might not be reflective of sgRNA effectiveness, or 2) 67.4% of the sgRNA hits did not induce target gene activation. To distinguish between these two possibilities, I analysed target gene expression as the average in all the cells expressing a given sgRNA (Appendix C, see Materials and Methods) and found that target gene activation could be detected in 26 out of these 46 sgRNA hits (56.5%) when considering a mean log fold change to non-targeting sgRNA controls > 0 (Figure 5.14) (Appendix C). Therefore, these results indicate that target gene activation in the population of cells expressing a given sgRNA can be induced even with less than 10 cells showing stringent target gene activation. As discussed in chapter 4, lack of detection of target gene activation for the remaining 20 sgRNA hits (43.5%) could be due to technical reasons such as drop-outs commonly associated to scRNA-seq data

or due to biological uncoupling between target gene upregulation and downstream transcriptional responses. Alternatively, it is also possible that lack of target gene activation in sgRNA hits is associated with false positives introduced during hit calling. In these regards, except for *Dppa2* and *Smarca5*, only one out of the two targeting sgRNAs included for the remaining 42 hits were ranked for inducing a ZGA-like signature (Figure 5.10). Strikingly, while for the top-ranking hits the sgRNA that induced ZGA-like expression (sgRNA hit) corresponded to the one triggering higher target gene activation out of the two, this was not always the case for lower-ranking hits (Figure 5.14). Since no major off-target effects were detected for these sgRNAs (Appendix C), this likely reflects noise in the hit calling performed based on MOFA ZGA-like variance and highlights the importance to perform screen validations with alternative methods and systems.

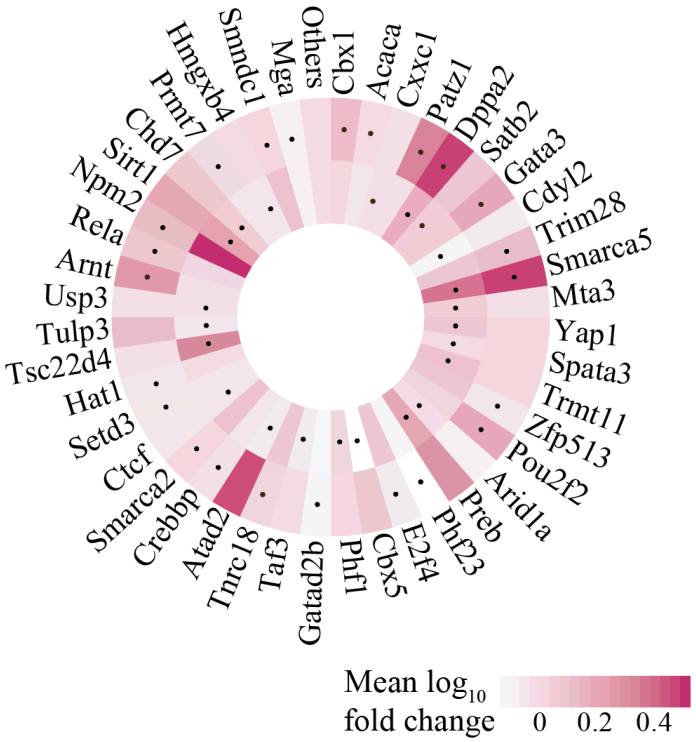



**Figure 5.14– Target gene expression in cells expressing sgRNA screen hits**

Target gene expression for the 46 sgRNA hits, measured by mean $\log_{10}$ fold change expression between SAM22 mESCs expressing the relevant targeting sgRNA and SAM22 mESCs expressing non-targeting sgRNA controls. Both sgRNAs targeting the gene of interest are shown (outer and inner circle) and sgRNAs that were identified as a hit for inducing a ZGA-like signature (as ranked by the contribution to MOFA factor 3) are highlighted with a black dot. Log fold change target gene expression values for every sgRNA and the complete sgRNA hit rank are available is Appendix C.

To validate the sgRNA hit calling by MOFA, I performed more conventional differential gene expression analysis. However, in most cases, only a small subset of genes was significantly differentially expressed (EdgeR, FDR<10%) in SAM22 mESCs expressing sgRNA hits compared to cells expressing non-targeting sgRNA controls (between 0 and 224 genes for different sgRNA hits, median 4.5) (Appendix C and see Materials and Methods), consistent with the sparsity observed in 10X Genomics scRNA-seq data. To overcome this limitation, I considered the top 400 upregulated genes for each sgRNA, ranked them by statistical significance and then analysed the cumulative frequency of known ZGA genes, as described in published literature (Eckersley-Maslin et al. 2016; Hendrickson et al. 2017; Y. Li et al. 2018) (Appendix I), in this rank. This analysis revealed that 32 out of the 46 sgRNA hits (69.6%) induced a transcriptome downstream of them that was prominently enriched for known ZGA transcripts, compared to the background level of enrichment calculated by differential gene expression between cells expressing non-targeting sgRNA controls (Figure 5.15) (Appendix C). Importantly, the remaining 14 sgRNA hits called by MOFA that overlapped with non-targeting sgRNA controls in this differential gene expression rank (Figure 5.15) were not associated with lower effects on target gene activation (Figure 5.16) or with a lower position in the MOFA rank (Figure 5.10, Appendix C). Therefore, I can speculate about three reasons that could explain why these 14 sgRNA hits were not detected as statistically significant in the differential gene expression ZGA rank: 1) they are strong inducers of ZGA-associated repeat elements but not as much of ZGA genes and, consequently, they were captured by MOFA but not by this rank which only considers protein-coding genes; 2) lack of power to detect the effects on individual genes in a pre-defined list of ZGA genes (as defined in Appendix I) might have hampered the identification of relevant hits using differential gene expression analysis, while they could be identified by MOFA because it considers a gene signature rather than individual genes; or 3) the hit calling performed based on MOFA introduced false positive hits. Either way, these results validate the MOFA analysis for the majority of screen hits.

In summary, using MOFA to integrate the expression of protein-coding genes and repeat elements in the CRISPRa scRNA-seq dataset, we identified 44 factors whose activation induced a ZGA-like transcriptional response. These included four positive controls known to regulate ZGA, namely *Dppa2* (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Y.-L. Yan et

al. 2019), *Yap1* (C. Yu, Ji, Dang, et al. 2016), *Ctcf* (Wan et al. 2008) and *Gata3* (Eckersley-Maslin et al. 2019). Interestingly, two of the candidates included as negative controls for not being expressed in oocytes or during pre-implantation development, *Spata3* and *Gpr6*, were obtained as top-ranking hits (Figures 5.10 and 5.15), indicating they might have a function in inducing 2C-like mESCs but could not regulate ZGA *in vivo*. These observations emphasise the need to validate these hits in the embryo.

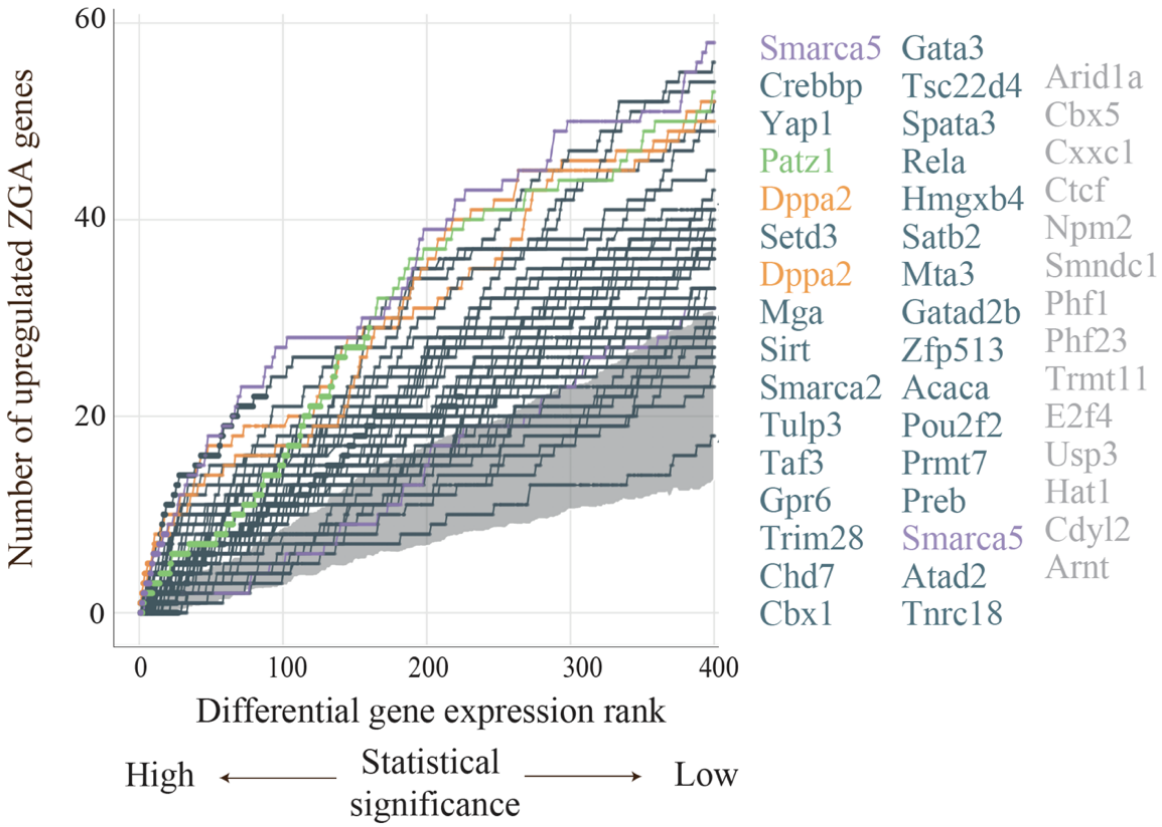



**Figure 5.15– Differential gene expression ZGA rank for sgRNA screen hits**

Cumulative rank of the number of previously described ZGA genes (as described in Appendix I - Eckersley-Maslin et al. 2016; Hendrickson et al. 2017; Y. Li et al. 2018) upregulated in SAM22 mESCs expressing each sgRNA hit compared to non-targeting sgRNA controls, considering the top 400 genes ranked by statistical significance of differential gene expression test (generalised linear model likelihood ratio test as implemented in EdgeR, FDR<100%). The background distribution based on differential gene expression between SAM22 mESCs expressing non-targeting sgRNA controls is shown in grey, displaying plus and minus one standard deviation around the mean of ZGA genes recovered by non-targeting sgRNAs. The names of the target genes for sgRNAs hits are depicted from top to bottom and left to right according to the rank order based on number of ZGA genes recovered, with those for which the differential gene expression rank overlaps with the non-targeting sgRNA control background shown in grey in the third column. The top-ranking sgRNA hits for *Dppa2* (orange), *Smarca5* (purple) and *Patz1* (green) are highlighted. The number of upregulated ZGA genes recovered for each sgRNA is available in Appendix C.

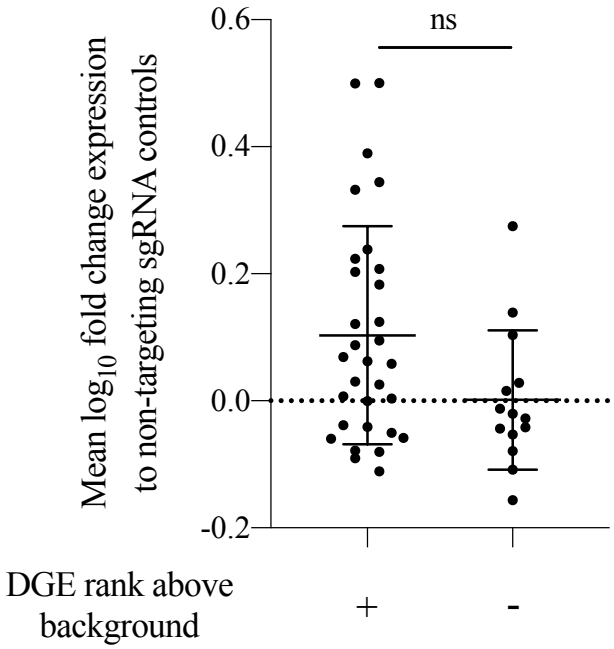



**Figure 5.16– Target gene expression in cells expressing sgRNA screen hits, classified according to the differential gene expression ZGA rank**

Target gene expression for the 46 sgRNA hits, measured by mean $\log_{10}$ fold change expression between SAM22 mESCs expressing the relevant targeting sgRNA and SAM22 mESCs expressing non-targeting sgRNA controls, grouped according to their performance in the differential gene expression rank ZGA rank (DGE rank). Differences between the two groups assessed by Mann-Whitney test were not statistically significant. The classification of the 46 sgRNA hits according to their performance in the DGE rank is based on Figure 5.15 and available in Appendix C.

### 5.2.3 Validation of *Dppa2*, *Smarca5* and *Patz1* as regulators of ZGA in mESCs

Next, I selected three top-ranking screen hits to validate *in vitro* in mESCs. As discussed earlier, false positive and negative hits can arise from pooled screens and, therefore, individual validations of hits using CRISPRa and alternative methods are necessary before considering their molecular mechanism of action. I validated the DNA binding protein *Dppa2*, the ATPase subunit of the ISWI chromatin remodeling complex *Smarca5* and the transcription factor *Patz1* (Figure 5.17). Both *Dppa2* and *Smarca5* were strong hits because both of their targeting sgRNAs induced a ZGA-like response in the screen (Figures 5.10 and 5.15). Moreover, while this study was undergoing, *Dppa2* was confirmed in the Reik laboratory and by other groups as a key regulator of ZGA-like transcriptional programmes using cDNA overexpression and KO approaches in mESCs (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Y.-L. Yan et al. 2019). Although not comprehensively studied for its role in ZGA, *Smarca5* has been previously implicated in early embryogenesis with hints to a role in transcriptional regulation (Stopka & Skoultchi 2011; Torres-Padilla & Zernicka-Goetz 2006). Therefore, the identification of *Dppa2* and *Smarca5* in the screen supports my pooled CRISPRa approach coupled to scRNA-seq as a reliable method to identify relevant regulators. Excitingly, the top-ranking hit transcription factor *Patz1* (Figures 5.10 and 5.15), although previously linked to mESC pluripotency (Ow et al. 2014), had never been studied in the context of ZGA. In this validation panel, I also included the negative control candidate *Carhsp1* (Figure 5.17), since both of its targeting sgRNAs induced effective target gene activation without triggering ZGA-like transcription (Appendix C). Other candidate genes were also validated as described in Appendix J.

I used two alternative methods of gene overexpression to validate these screen candidates (Figure 5.17). Firstly, CRISPRa by transducing one of the sgRNAs used in the screen (Appendix C) cloned into the CROP-sgRNA-MS2 construct into SAM22 mESCs, followed by antibiotic selection and cell expansion for 10 days; and secondly, ectopic cDNA overexpression by transiently transfecting a cDNA-eGFP fusion construct into E14 mESCs for 48 hours, followed by FACS sorting of $GFP^+$ cells. In both cases, the transcriptome was analysed by bulk

poly-A capture RNA-sequencing (Figure 5.17) and both set of experiments were performed in biological triplicate.

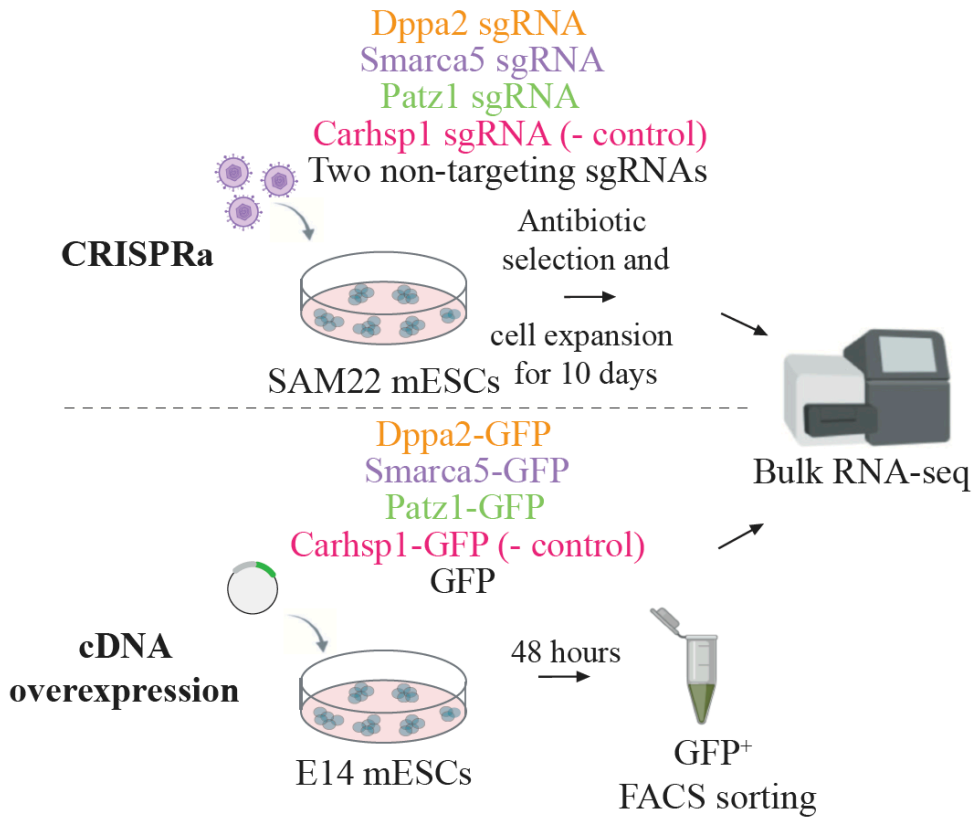



**Figure 5.17– Approaches for hit validation: CRISPRa and cDNA overexpression**

Schematic representation of two sets of validation approaches to confirm the screen hits *Dppa2* (orange), S*marca5* (purple) and *Patz1* (green), using *Carhsp1* (pink) as a negative control candidate. The first approach consists of lentiviral transductions (purple viral particles) with individual effective sgRNAs targeting gene promoters and cloned into the CROP-sgRNA-MS2 construct into SAM22 mESCs, followed by antibiotic selection, cell expansion for 10 days and analysis by bulk RNA-sequencing (top panel). The second approach consist of ectopically overexpressing the gene of interest by cloning its cDNA into an eGFP fusion construct and transiently transfecting it into E14 mESCs, followed by fluorescence-activated cell sorting (FACS) of $GFP^+$ cells 48 hours later and analysis by bulk RNA-sequencing (bottom panel). Control samples were lentiviral transductions of two non-targeting sgRNAs into SAM22 mESCs for CRISPRa (top panel) and $GFP^+$-only transfection in E14 mESCs for cDNA overexpression (bottom panel). Both sets of experiments were done in triplicate. The sgRNAs used for CRISPRa in these bulk RNA-sequencing experiments are described in Appendix C.

The sgRNAs selected for validation experiments had shown to be effective in triggering target gene activation in the scRNA-seq CRISPRa screen (Figure 5.18). Consistently, both individual CRISPRa and cDNA transient transfection led to target gene upregulation, as analysed by bulk RNA-sequencing (Figure 5.19).

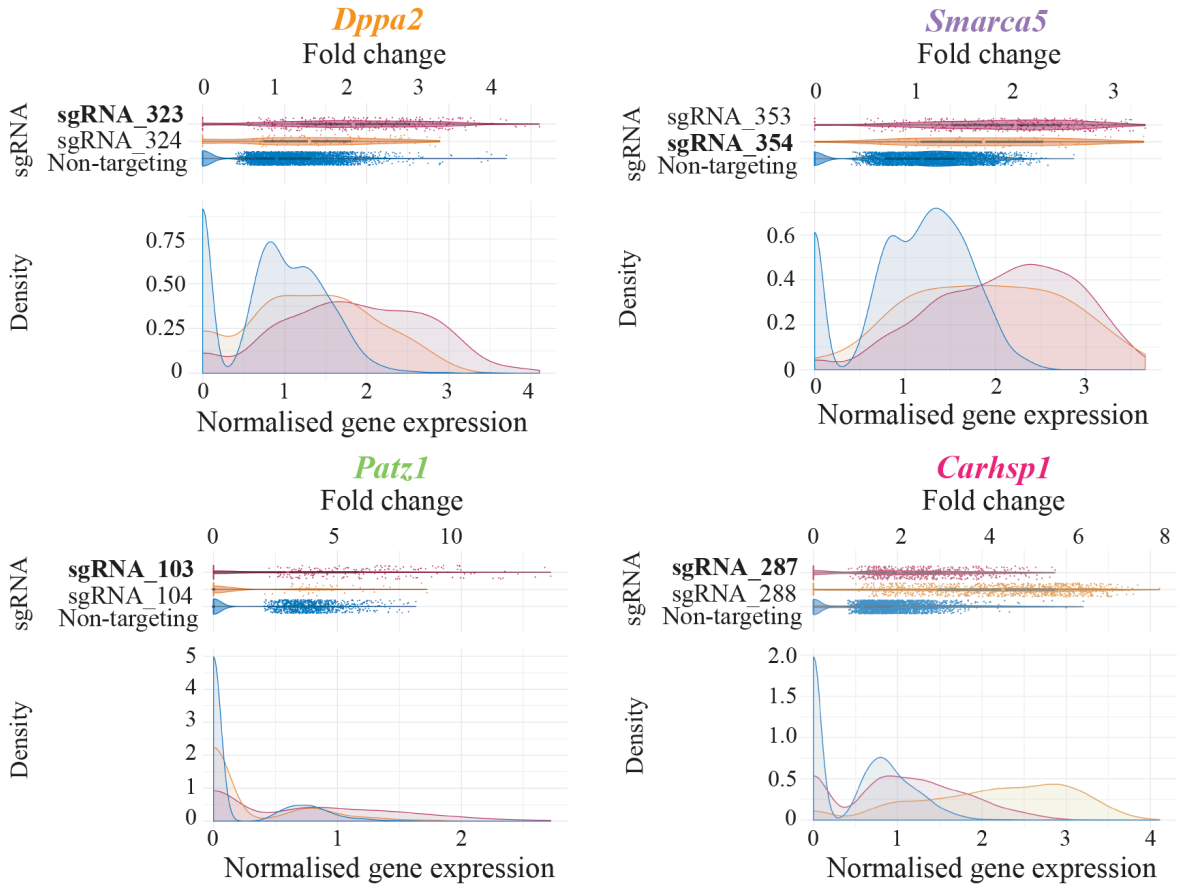



**Figure 5.18– Target gene expression of *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1* in the CRISPRa screen dataset analysed by scRNA-seq**

Dot-plot (top panels) and density representation (bottom panels) of expression levels induced by CRISPRa of *Dppa2* (top left, orange), *Smarca5* (top right, purple), *Patz1* (bottom left, green) and *Carhsp1* (bottom right, pink) in SAM22 mESCs expressing their respective targeting sgRNAs (red and orange, see Appendix C for sgRNA identifiers) and in SAM22 mESCs expressing either of the 15 non-targeting sgRNA controls (blue). In the dot-plots, each dot represents a cell plotted as a function of fold change expression to the average expression in the non-targeting sgRNA control cells. In the density plots, normalised gene expression is calculated as $\log_e$(number of UMIs for target gene / sum (number of total UMIs in a cell) *10,000 + 1) and plotted as a function of density. The sgRNA for each target gene used for validation bulk RNA-sequencing experiments is highlighted in bold.

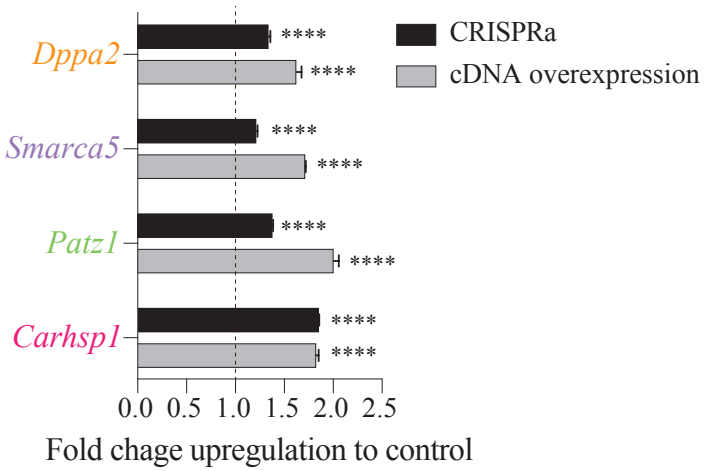



**Figure 5.19– Target gene expression of *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1* after CRISPRa and cDNA overexpression, analysed by bulk RNA-sequencing**

Fold change upregulation of *Dppa2* (orange), *Smarca5* (purple), *Patz1* (green) and *Carhsp1* (pink) by CRISPRa (black) and cDNA overexpression (grey) compared to respective controls (two non-targeting sgRNAs for CRISPRa and $GFP^+$-only for cDNA overexpression), measured by bulk RNA-sequencing. Data is shown as mean plus standard deviation of three biological replicates. Differences to controls are statistically significant (homoscedastic two-tailed t-test, ****: p-value < 0.0001). A dashed line at x = 1 is used to indicate the minimum value to observe fold change upregulation to controls. The sgRNAs used for CRISPRa in these bulk RNA-sequencing experiments are described in Appendix C.

Next, I assessed the similarities between the transcriptional changes captured by scRNA-seq and bulk RNA-sequencing upon CRISPRa of these targets and, overall, I found that despite the increased power in calling differential gene expression using bulk RNA-sequencing, they were highly correlated (Figure 5.20). Then, similar to the comparison analysis between CRISPRa and cDNA overexpression presented in chapter 3 (section 3.2.2), I calculated correlation coefficients between transcriptome-wide bulk gene expression profiles of *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1* CRISPRa and cDNA overexpression (Figure 5.21). Interestingly, I observed markedly correlated transcriptional responses between CRISPRa and cDNA overexpression of the same target, but also strikingly similar profiles induced by *Dppa2*, *Smarca5* and *Patz1* by either overexpression method (Figure 5.21). Furthermore, the transcriptome induced by these three screen hits was remarkably distinct from the transcriptome triggered by the negative control *Carhsp1* (Figure 5.21), consistent with the small effects that *Carhsp1* CRISPRa had on mESCs (Figure 5.20). Together, these validation experiments confirmed that CRISPRa coupled with scRNA-seq readout is a robust method to assess the transcriptional responses triggered by gene overexpression.

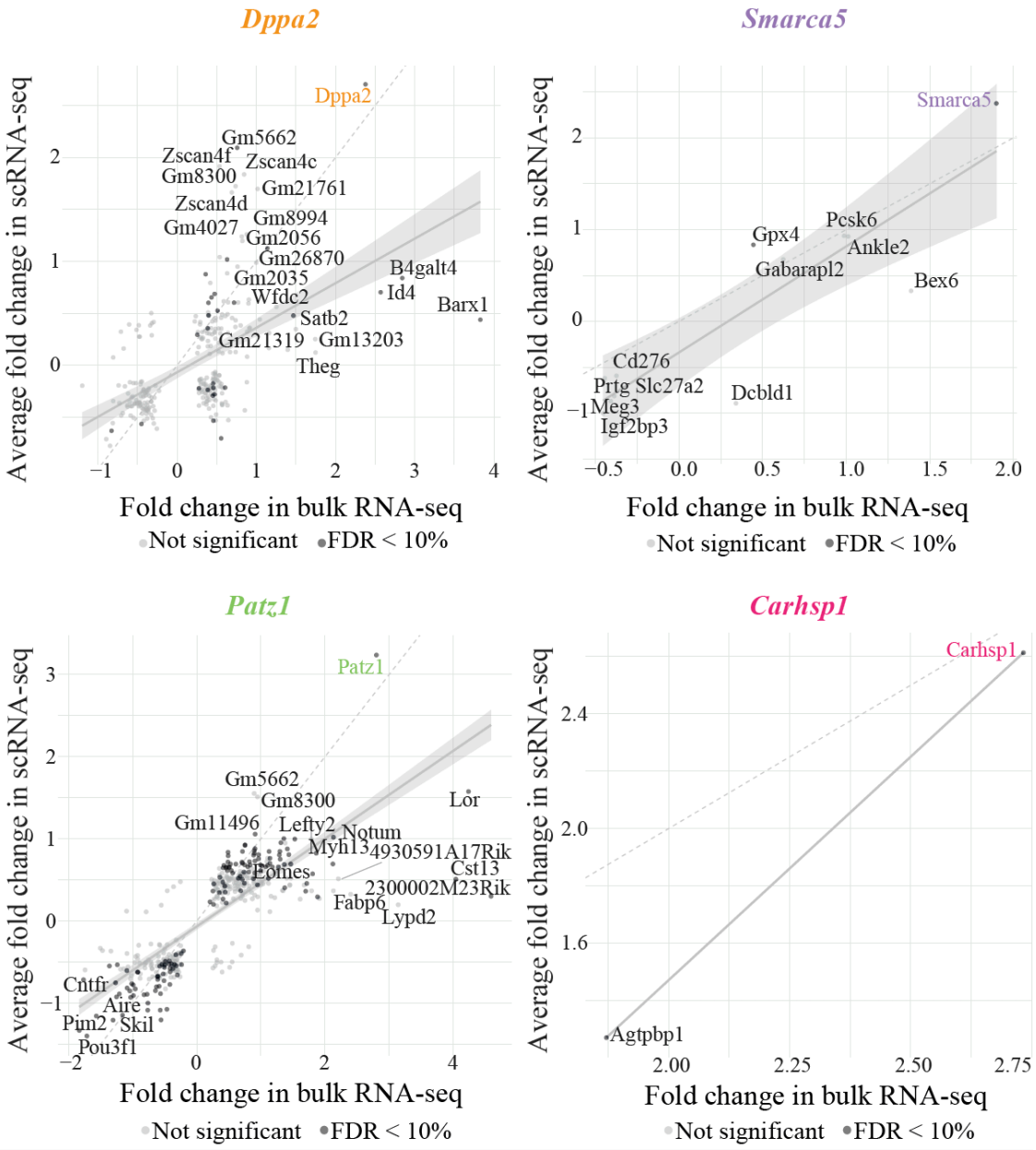



**Figure 5.20– Correlation between the transcriptional changes captured by scRNA-seq and bulk RNA-sequencing after CRISPRa of *Dppa2*, *Smarca5*, Patz1 and *Carhsp1***

Scatterplot between fold change values of differentially expressed genes estimated based on bulk CRISPRa RNA-sequencing data (EdgeR, FDR<10%) (x axis) and fold change estimates for the corresponding genes in cells expressing the same sgRNA based on the CRISPRa scRNA-seq data (y axis) for the target genes *Dppa2* (orange, top left), *Smarca5* (purple, top right), *Patz1* (green, bottom left) and *Carhsp1* (pink, bottom right). Genes that were also differentially expressed in scRNA-seq data (EdgeR, FDR<10%) are labelled in dark grey whereas those genes differentially expressed by bulk RNA-sequencing but not in scRNA-seq are labelled in light grey (not significant). In each panel, the top 20 genes with the highest average log fold change in scRNA-seq and bulk RNA-sequencing are labelled and a regression line was fitted to highlight the correlation trend between both read-outs. Dashed lines mark a $y = x$ line. The sgRNAs used for CRISPRa in bulk RNA-sequencing experiments are described in Appendix C.

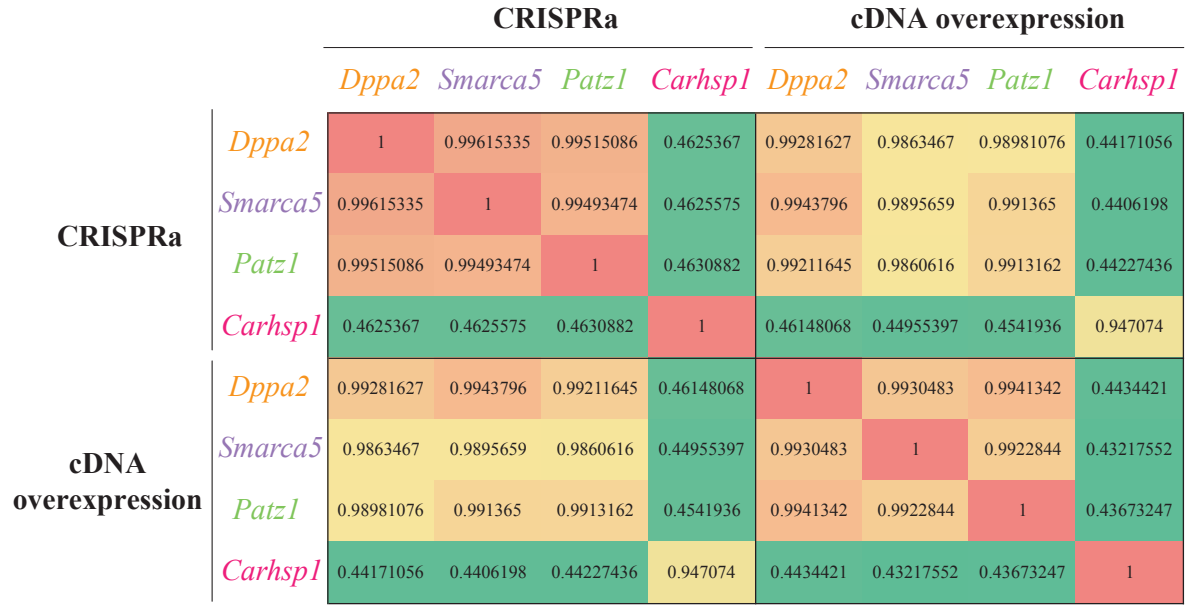



**Figure 5.21– Correlation heatmap between CRISPRa and cDNA overexpression of *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1*, as analysed by bulk RNA-sequencing**

Heatmap of Pearson correlation coefficients (displayed in numbers in each square) between bulk RNA-sequencing gene expression profiles of *Dppa2* (orange), *Smarca5* (purple), *Patz1* (green) and *Carhsp1* (pink) CRISPRa and cDNA overexpression. The scale from green to red represents low to high correlation. The sgRNAs used for CRISPRa in these bulk RNA-sequencing experiments are described in Appendix C.

The strikingly similar transcriptional responses observed upon upregulation of *Dppa2*, *Smarca5* and *Patz1* (Figure 5.21) suggested they might regulate similar transcriptional networks. To validate whether these networks corresponded to a ZGA-like response, I analysed the expression levels of the genes capture by the MOFA ZGA-like factor in these bulk RNA-sequencing validation experiments and observed that both CRISPRa and cDNA overexpression of *Dppa2*, *Smarca5* and *Patz1*, but not *Carhsp1*, induced a ZGA-like response (Figure 5.22). Importantly, despite the high correlation between the transcriptome-wide changes triggered by CRISPRa and cDNA overexpression of these hits (Figure 5.21), the ZGA-like transcriptome induced presented differences both between target genes and between overexpression methods (Figure 5.22). Therefore, these results suggest that *Dppa2*, *Smarca5* and *Patz1* might regulate different networks within the ZGA programme and, additionally, the differences in experimental design (Figure 5.17) and target gene dosage (Figure 5.19) could explain the slightly different gene expression profiles triggered by either method (Figure 5.21).

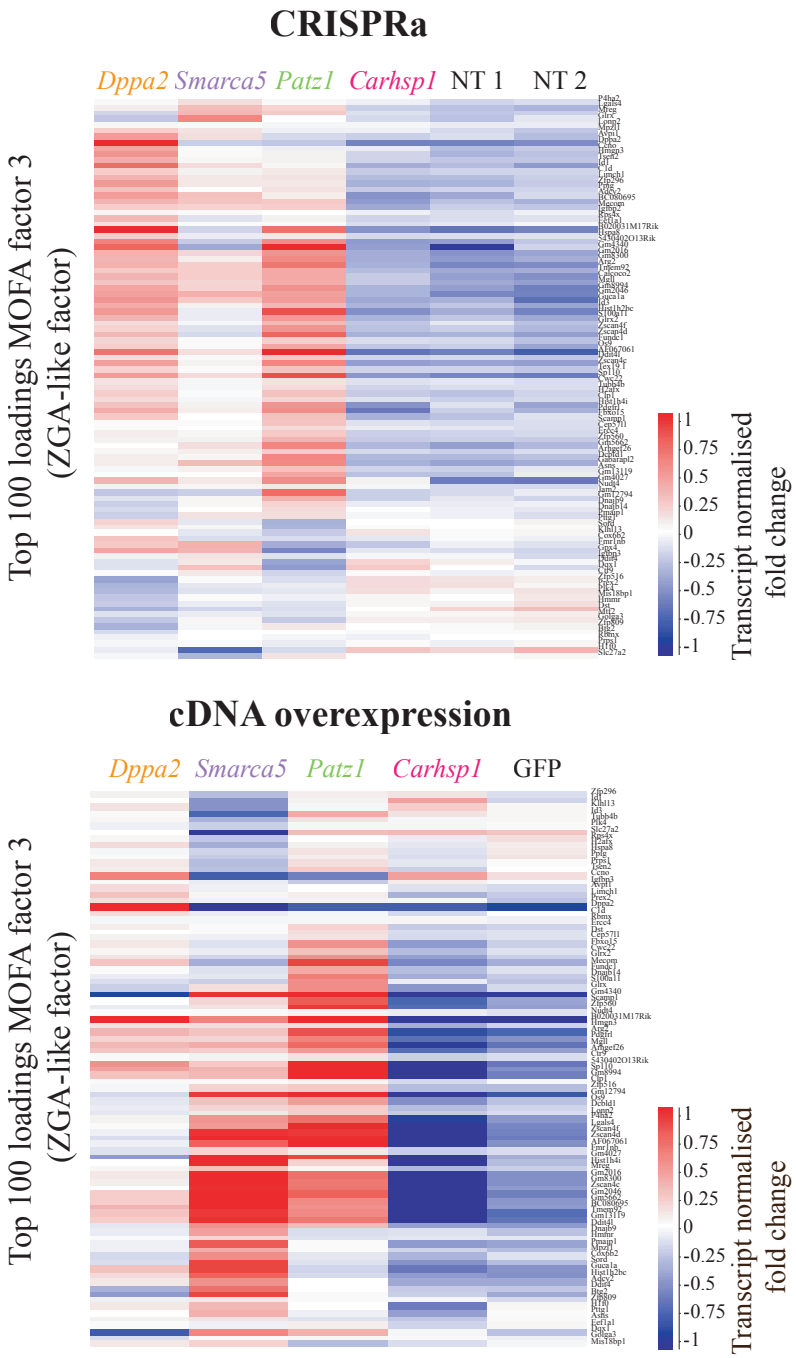



**Figure 5.22– Heatmaps of ZGA-like gene expression after CRISPRa and cDNA overexpression of *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1*, analysed by bulk RNA-sequencing**

Heatmaps showing transcript normalised expression, scaled per transcript, of the top 100 loading genes for MOFA factor 3 (ZGA-like factor), in bulk RNA-sequencing libraries of *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1* CRISPRa (top panel) and cDNA overexpression (bottom panel). Controls are two different non-targeting sgRNAs (NT1 and NT2) for CRISPRa and GFP only overexpression for cDNA. The sgRNAs used for CRISPRa in these bulk RNA-sequencing experiments are described in Appendix C.

Furthermore, MERVL and ERV1 elements, which captured significant variability amongst the repeat families in the scRNA-seq dataset as analysed by MOFA (Figure 5.9), were significantly upregulated by *Dppa2*, *Smarca5* and *Patz1*, but not by *Carhsp1*, regardless of the overexpression method (Figure 5.23). Major satellites, despite ranking second top of the repeat families in the variability explained by the MOFA ZGA-like factor (Figure 5.9) and being overall upregulated by SAM22 mESCs expressing the 46 sgRNA hits compared to cells expressing other sgRNAs (Figure 5.12), were not specifically upregulated by *Dppa2*, *Smarca5*, *Patz1* or *Carhsp1*. SINE B2 elements, also expressed during the major wave of ZGA (Vasseur et al. 1985) were slightly induced by *Dppa2*, *Smarca5* and *Patz1*, although not consistently between CRISPRa and cDNA overexpression methods. LINE-1 expression, also linked to ZGA (Percharde et al. 2018) was consistently upregulated by *Dppa2* and *Patz1*, but not by *Smarca5* (Figure 5.23). The rest of the repeat families analysed (ERVK, SINE Alu B1 and SINE B4) were overall not induced upon overexpression of these factors (Figure 5.23), consistent with the low contribution of these families to the MOFA ZGA-like factor variance (Figure 5.9) and the lack of upregulation by the 46 sgRNA screen hits (Figure 5.12).

Lastly, to account for the differences between CRISPRa and cDNA overexpression (Figure 5.22), I defined highly-confident differentially expressed genes downstream of *Dppa2*, *Smarca5* and *Patz1* as those differentially expressed both after CRISPRa and cDNA overexpression (intersect of EdgeR FDR <5% to respective controls). Notably, highly-confident upregulated genes by *Dppa2*, *Smarca5* and *Patz1* showed a peak of expression in mid-to-late two-cell embryos, compared to a random set of expressed genes in the bulk RNA-sequencing dataset (Figure 5.24), suggesting these maternal factors regulate genes expressed during the major wave of ZGA. In conclusion, these validation approaches provided robust evidence that *Dppa2*, *Smarca5* and *Patz1* activate a ZGA-like programme in mESCs and are, therefore, strong candidates for ZGA regulation *in vivo*.

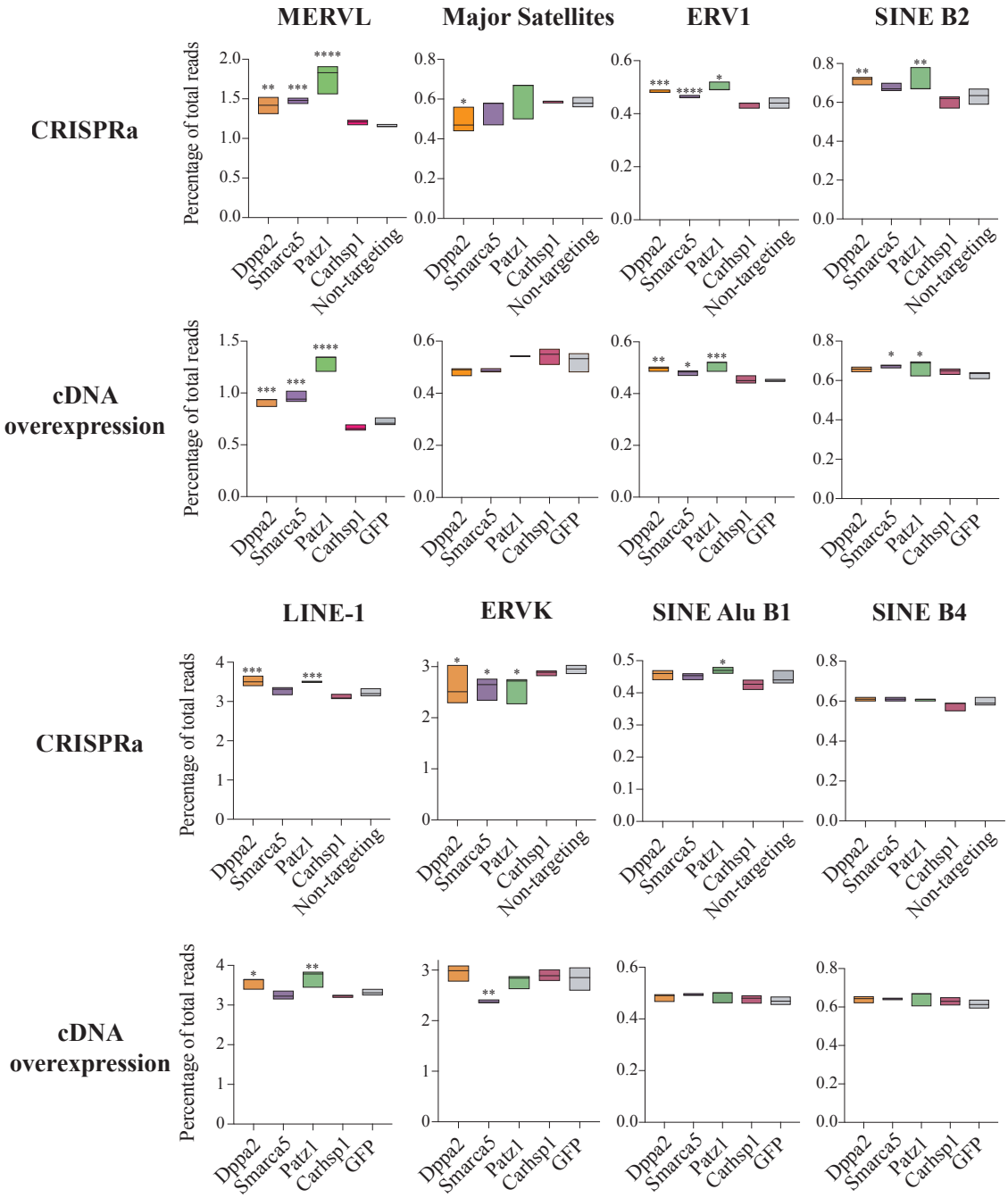



**Figure 5.23– Expression of repeat elements upon CRISPRa and cDNA overexpression of *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1*, analysed by bulk RNA-sequencing**

Box-whisker plots of three biological replicates showing expression of different repeat families in percentage of total reads, measured by bulk RNA-sequencing after CRISPRa and cDNA overexpression of *Dppa2* (orange), *Smarca5* (purple), *Patz1* (green) and *Carhsp1* (pink) and in controls (average of two non-targeting sgRNAs for CRISPRa and GFP$^+$-only for cDNA overexpression, grey). Statistically significant differences to controls are reported as ****: p-value < 0.0001, ***: p-value <0.001, **: p-value <0.01, *: p-value <0.5, absence of stars: non-significant homoscedastic two-tailed t-test. The sgRNAs used for CRISPRa in these bulk RNA-sequencing experiments are described in Appendix C.

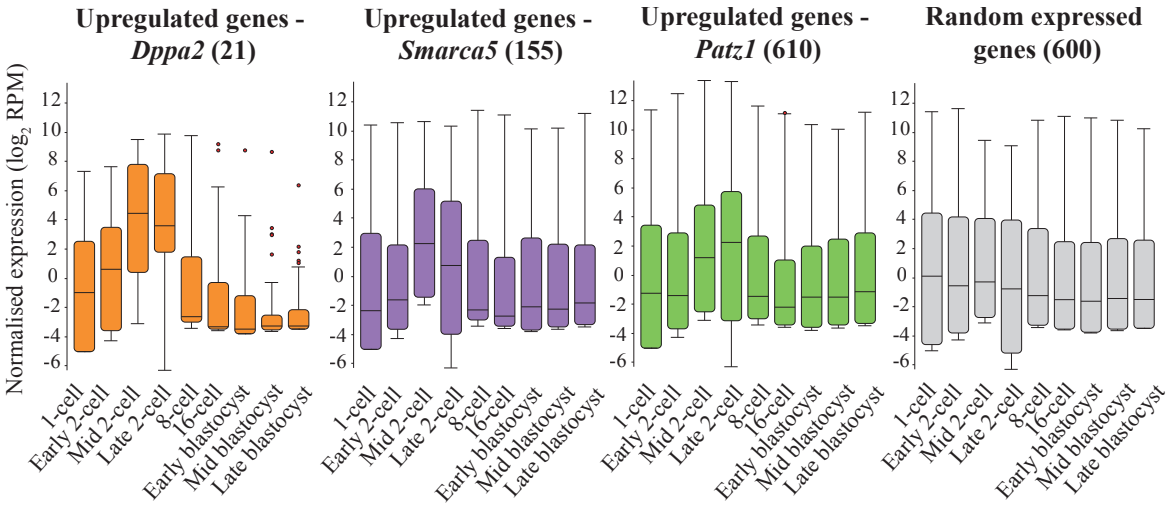



**Figure 5.24– Expression of upregulated genes by *Dppa2*, *Smarca5* and *Patz1* in pre-implantation development**

Box-whisker plots showing normalised expression in $\log_2$ reads per million (RPM) of differentially upregulated genes by both CRISPRa and cDNA overexpression of *Dppa2* (orange), *Smarca5* (purple), *Patz1* (green), and *Carhsp1* (pink), as well as of a random set of expressed genes (grey) in the bulk RNA-sequencing dataset, during mouse pre-implantation development (data analysed from Deng et al. 2014). Differential gene expression was calculated by EdgeR (FDR <5%). The number of analysed genes in each case is depicted in brackets and corresponds to the intersect of differentially upregulated genes by both CRISPRa and cDNA overexpression of each target. The sgRNAs used for CRISPRa in these bulk RNA-sequencing experiments are described in Appendix C.

### 5.2.4 Interplay between *Smarca5* and *Dppa2* during ZGA regulation

I next focused on the mechanistic regulation of the validated screen hits *Dppa2* and *Smarca5*. I chose to follow up the molecular mechanism of these two hits based on a previously proposed hypothesis that DPPA2 protein might require an open chromatin landscape to regulate mESC entry to the 2C-like state (Eckersley-Maslin et al. 2019). This hypothesis is derived from the observation that DPPA2 is expressed across all mESCs but it only induces ZGA-like transcription in a subset of them, possibly those with chromatin permissive conditions (Eckersley-Maslin et al. 2019). With this in mind, I hypothesised that SMARCA5, as a strong regulator of ZGA (Figures 5.10, 5.15, 5.22-5.24), might help providing such chromatin conditions to regulate DPPA2's function during ZGA.

Since SMARCA5 is the ATPase subunit of a larger protein complex, the ISWI chromatin remodeling complex (reviewed in Hota & Bruneau 2016), I first analysed recently published

RNA-sequencing data (Barisic et al. 2019) to investigate whether SMARCA5 exerts its ZGA regulatory function through its catalytic activity or through interactions with accessory subunits of the ISWI complex. In this published dataset, the authors generated bulk RNA-sequencing libraries of *Smarca5* KO mESCs and subsequent complementation of this cell line with either wild-type (WT) SMARCA5 or a catalytic mutant (Barisic et al. 2019). Remarkedly, I found 391 ZGA genes (as described in Appendix I) differentially downregulated in *Smarca5* KO mESCs compared to WT mESCs (Figure 5.25). The expression levels of most of these genes were partially or completely restored with the WT but not with the catalytically-dead version of SMARCA5 (Figure 5.25), indicating that the regulation of ZGA by SMARCA5 is dependent on its ATPase activity.

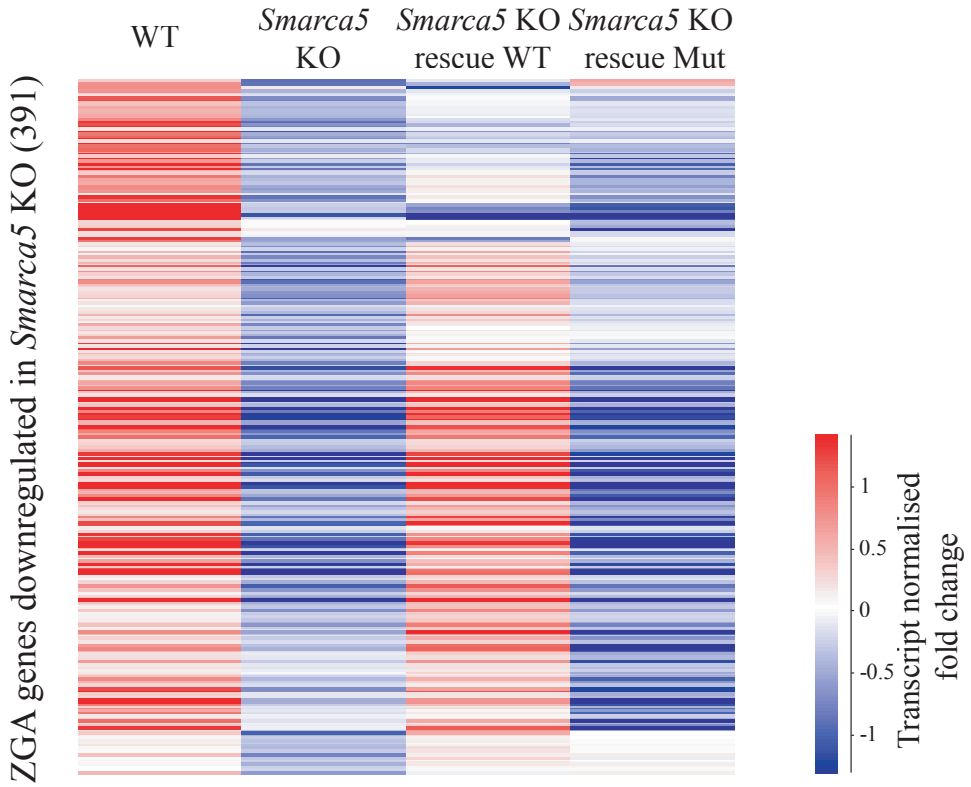



**Figure 5.25– Expression of ZGA genes in *Smarca5* KO mESCs rescued with wild-type or catalytically-dead mutant SMARCA5**

Heatmap showing normalised transcript expression, scaled per transcript, of downregulated ZGA genes in *Smarca5* knock-out (KO) mESCs compared to wild-type (WT) (EdgeR, FDR<5%, overlapping with the gene list from Appendix I), in WT mESCs, *Smarca5* KO mESCs and *Smarca5* KO mESCs expressing a SMARCA5 WT protein or a SMARCA5 catalytic-dead mutant protein (Mut) (data analysed from Barisic et al. 2019).

At the transcript level, both *Dppa2* and *Smarca5* are expressed in oocytes and throughout pre-implantation development (Figure 5.26). *Smarca5* is slightly more highly expressed in the oocyte than *Dppa2*, however, after fertilisation, *Dppa2* rapidly increases in its expression, specially after the major wave of ZGA at the two-cell stage (Figure 5.26), consistent with *Dppa2* being not only a maternal protein but also a transcript expressed from the zygotic genome during ZGA (Eckersley-Maslin et al. 2019; Iaco et al. 2019; Appendix I). *Smarca5* expression also increases at the time of major ZGA, which could indicate transcription from the zygotic genome, although to a lesser extent than *Dppa2* (Figure 5.26).

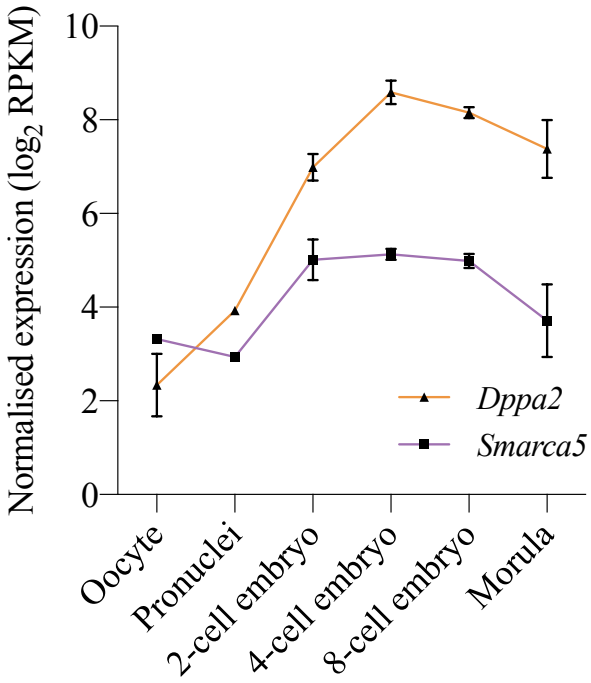



**Figure 5.26– Expression of *Dppa2* and *Smarca5* during pre-implantation development**

Normalised expression in $log_2$ reads per kilobase per million (RPKM) of *Dppa2* (orange triangles) and *Smarca5* (purple squares) in mouse oocytes and pre-implantation development (data analysed from Xue et al. 2013). Data is shown as mean plus and minus standard deviation of 2-3 biological replicates.

At the protein level, immunofluorescence analysis performed by Oana Kubinyecz and Fátima Santos in the Reik laboratory show that DPPA2 localised in the cytoplasm of zygotes, while SMARCA5 was present in both pronuclei (Figure 5.27). Strikingly, in two-cell embryos, at the time of the major wave of ZGA, DPPA2 was found in the nuclei co-localising with SMARCA5 (Figures 5.27 and 5.28). Given the uncoupling between transcription and translation in early embryos (Nothias et al. 1996; Gao et al. 2019), these results suggest translocation of the maternal DPPA2 protein from the cytoplasm to the nuclei in the transition from zygote to two-cell stage, rather than new DPPA2 protein being produced from the zygotic genome. Importantly, a previous study has shown that DPPA2 and SMARCA5 physically interact in mESCs (Hernandez et al. 2018). Together, their co-localisation in two-cell embryos (Figures 5.27 and 5.28), physical interaction in mESCs (Hernandez et al. 2018) and similar transcriptional effects (Figures 5.21-5.24) suggest that DPPA2 and SMARCA5 act together to regulate their ZGA targets.

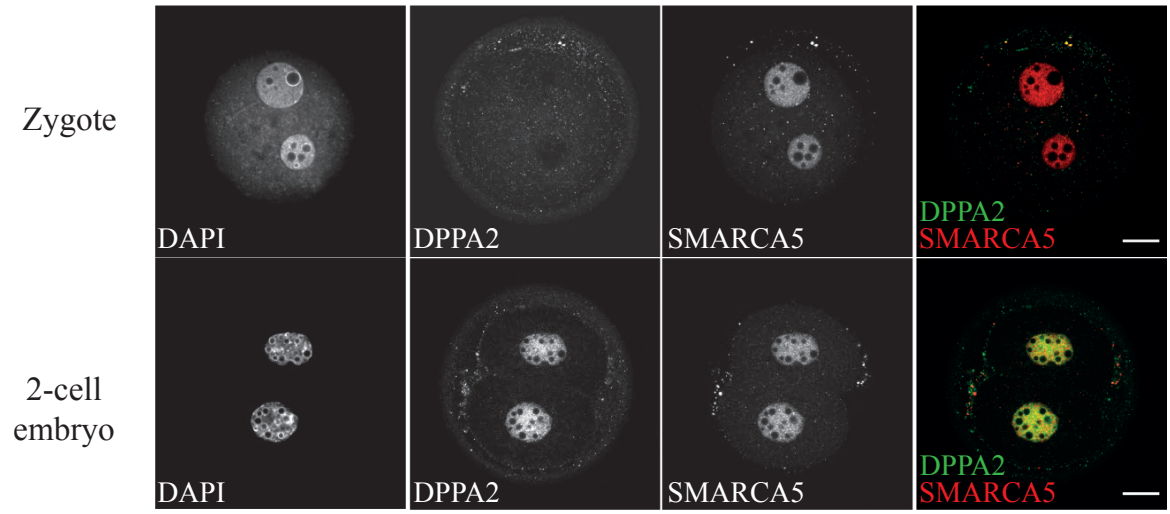



**Figure 5.27– Expression and localisation of DPPA2 and SMARCA5 proteins in zygotes and two-cell embryos**

Representative single optical slices of mouse zygotes (top row) and two-cell stage embryos (bottom row) immunostained for DPPA2 (green) and SMARCA5 (red). DNA was stained with DAPI to mark the pronuclei (zygote) and nuclei (two-cell embryos). Scale bars represent 20 μm.

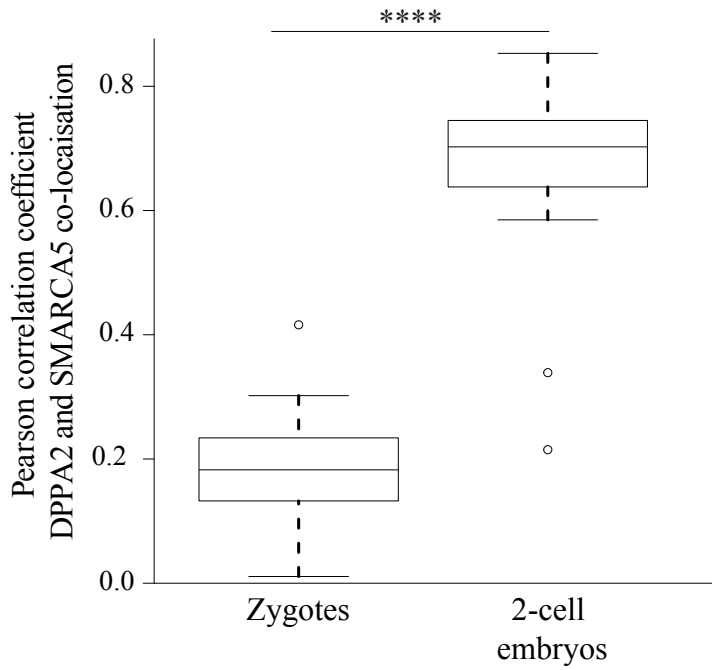



**Figure 5.28– Co-localisation of DPPA2 and SMARCA5 in zygotes and two-cell embryos**

Box-plots showing Pearson correlation coefficients calculated for co-localisation of DPPA2 and SMARCA5 in the pronuclei of 10 mouse zygotes and in the nuclei of 10 two-cell stage mouse embryos, analysed by immunofluorescence. Co-localisation values in the two pronuclei of zygotes and nuclei of each blastomere in two-cell embryos were measured separately and values were comparable. DPPA2 and SMARCA5 co-localise in two-cell embryos but not in zygotes (****: p-value < 0.0001; Mann-Whitney test).

Next, I investigated the mechanistic interdependencies, if any, between *Dppa2* and *Smarca5* through a series of KO and cDNA overexpression experiments analysed by qPCR. First, to test whether SMARCA5 protein is required for DPPA2's function, I overexpressed DPPA2-eGFP in WT and *Smarca5* KO mESCs (*Smarca5* KO mESCs from Barisic et al. 2019). After 48 hours transfection and $GFP^+$ FACS sorting, I first confirmed *Dppa2* overexpression and *Smarca5* KO at the transcript level, as well as maintenance of pluripotency measured by the marker *Pou5f1* (also known as *Oct4*) (Figure 5.29). Analysis of a panel of ZGA genes confirmed the RNA-sequencing results from published data (Barisic et al. 2019) (Figure 5.25) showing downregulation, although not complete loss, of these transcripts in *Smarca5* KO mESCs (Figure 5.30), in agreement also with the overexpression results in the screen and validation experiments (Figures 5.10, 5.15, 5.22-5.24). Excitingly, the expression of these ZGA genes was partially rescued in *Smarca5* KO mESCs after DPPA2 overexpression (Figure 5.30), suggesting SMARCA5 regulates ZGA transcription upstream of DPPA2 in mESCs.

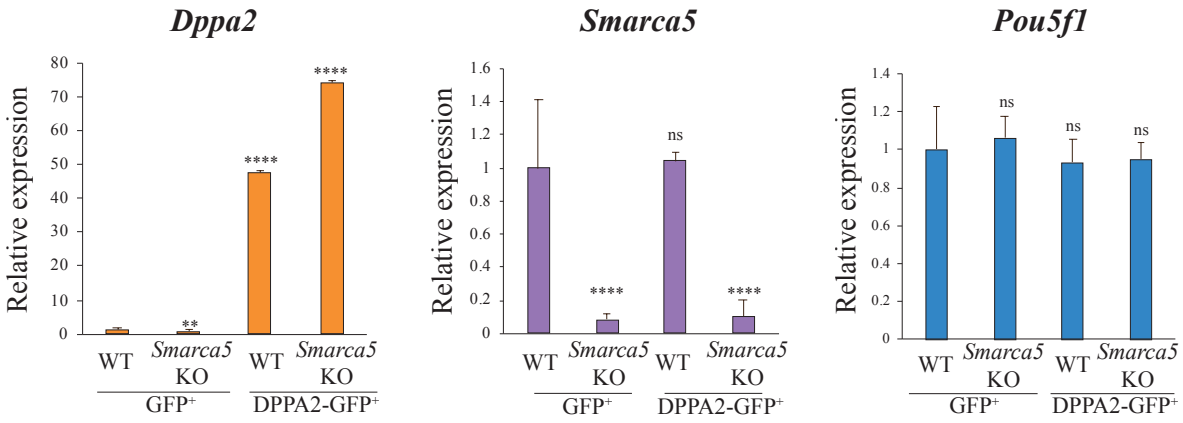



**Figure 5.29–*Dppa2*, *Smarca5* and *Pou5f1* expression in *Smarca5* KO mESCs transfected with DPPA2-GFP, analysed by qPCR**

Analysis of *Dppa2* (left panel, orange), *Smarca5* (middle panel, purple) and *Pou5f1* (*Oct4*) (right panel, blue) relative mRNA levels by qPCR in wild-type (WT) and *Smarca5* knock-out (KO) mESCs after 48 hours transient transfection of GFP or DPPA2-GFP and FACS-sorting for GFP$^+$ cells. The experiment was performed in triplicate and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to WT GFP$^+$ control are reported (homoscedastic two-tailed t-test, **: p-value <0.01, ****: p-value < 0.0001, ns: non-significant).

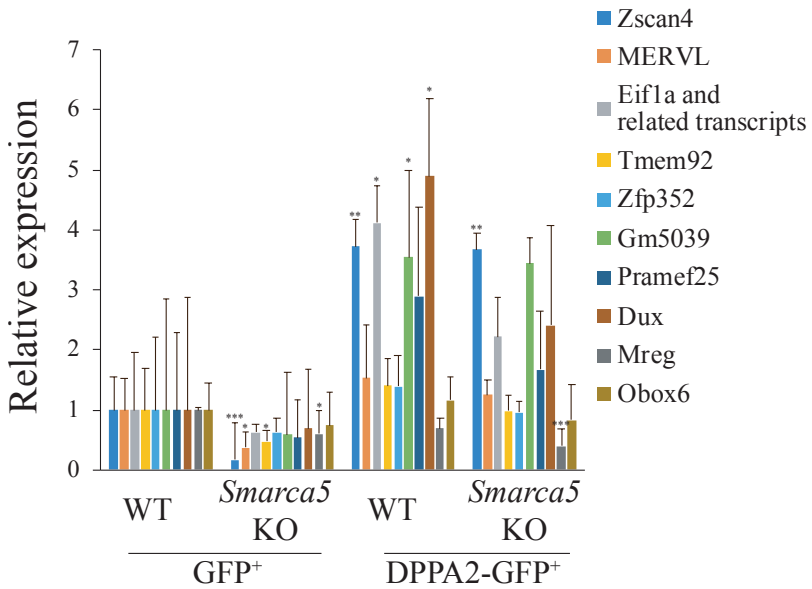



**Figure 5.30– ZGA markers expression in *Smarca5* KO mESCs transfected with DPPA2-GFP, analysed by qPCR**

Analysis of relative RNA levels of a panel of ZGA markers by qPCR in wild-type (WT) and *Smarca5* knock-out (KO) mESCs after 48 hours transient transfection of GFP or DPPA2-GFP and FACS-sorting for GFP$^+$ cells. The experiment was performed in triplicate and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to WT GFP$^+$ control are reported (homoscedastic two-tailed t-test, *: p-value <0.5, **: p-value <0.01, ***: p-value <0.001, absence of starts: non-significant).

To confirm these observations, I overexpressed SMARCA5-eGFP in WT and *Dppa2* KO mESCs (*Dppa2* KO mESCs from Eckersley-Maslin et al. 2019). Again, the transcript levels of *Smarca5* and *Dppa2* were as expected according to overexpression and KO phenotypes, and the pluripotency marker *Pou5f1* (also known as *Oct4*) remained largely unaltered (Figure 5.31). Importantly, although *Dppa2* transcript was not completely lost in *Dppa2* KO cells, its protein was undetectable by western-blot analysis (Appendix K, Eckersley-Maslin et al. 2019).

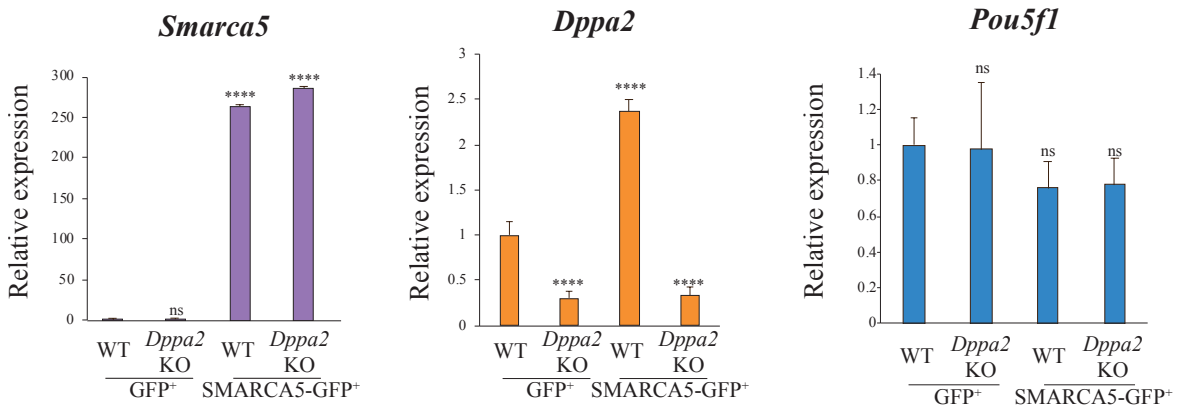



**Figure 5.31– *Smarca5*, *Dppa2* and *Pou5f1* expression in *Dppa2* KO mESCs cells transfected with SMARCA5-GFP, analysed by qPCR**

Analysis of *Smarca5* (left panel, purple), *Dppa2* (middle panel, orange) and *Pou5f1* (*Oct4*) (right panel, blue) relative mRNA levels by qPCR in wild-type (WT) and *Dppa2* knock-out (KO) mESCs after 48 hours transient transfection of GFP or SMARCA5-GFP and FACS-sorting for $GFP^+$ cells. The experiment was performed in triplicate and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to WT $GFP^+$ control are reported (homoscedastic two-tailed t-test, ****: p-value <0.0001, ns: non-significant).

Consistent with the screen results and validation experiments (Figures 5.10, 5.15, 5.22-5.24), in WT mESCs, SMARCA5 overexpression strongly upregulated ZGA genes, including *Dux*, the *Zscan4* cluster and MERVL, the former measured not only by qPCR but also by flow cytometry analysis of a fluorescent reporter incorporated in these cell lines (Figures 5.32 and 5.33). *Dppa2* KO mESCs presented nearly absent levels of ZGA transcripts (Figures 5.32 and 5.33), in agreement with the overexpression phenotype (Figures 5.10, 5.15, 5.22-5.24) and previous reports (Eckersley-Maslin et al. 2019). In these *Dppa2* KO mESCs, ZGA-like gene expression could not be rescued by SMARCA5 overexpression (Figures 5.32 and 5.33). These results show that DPPA2 is required for SMARCA5-mediated regulation of ZGA-like expression in mESCs, confirming SMARCA5 acts upstream of DPPA2.

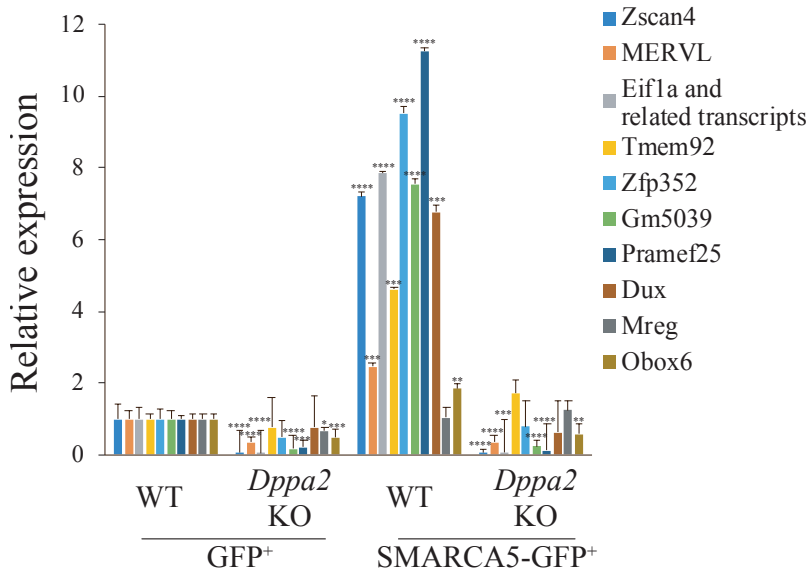



**Figure 5.32– ZGA markers expression in *Dppa2* KO mESCs transfected with SMARCA5-GFP, analysed by qPCR**

Analysis of relative RNA levels of a panel of ZGA markers by qPCR in wild-type (WT) and *Dppa2* knock-out (KO) mESCs after 48 hours transient transfection of GFP or SMARCA5-GFP and FACS-sorting for GFP$^+$ cells. The experiment was performed in triplicate and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to WT GFP$^+$ control are reported (homoscedastic two-tailed t-test, *: p-value <0.5, **: p-value <0.01, ***: p-value <0.001, ****: p-value <0.0001, absence of starts: non-significant).

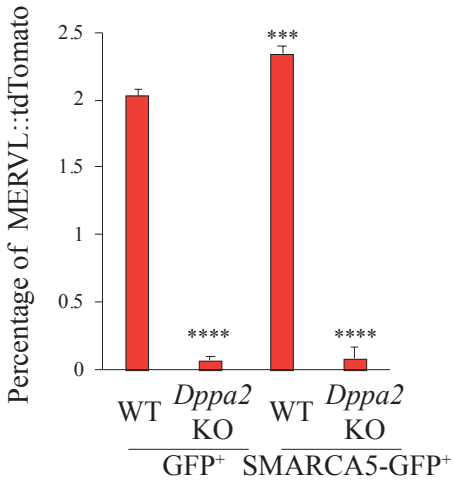



**Figure 5.33– MERVL reporter expression in *Dppa2* KO mESCs transfected with SMARCA5-GFP, analysed by flow cytometry**

Percentage of GFP$^+$ cells expressing a MERVL::tdTomato reporter in WT and *Dppa2* knock-out (KO) mESCs after 48 hours transient transfection of GFP or SMARCA5-GFP, analysed by flow cytometry. Data is shown as mean plus standard deviation of three biological replicates. Statistically significant differences to WT GFP$^+$ control are reported (homoscedastic two-tailed t-test, ***: p-value <0.001, ****: p-value <0.0001).

Notably, in WT cells, SMARCA5 overexpression resulted in *Dppa2* upregulation (Figure 5.31, middle panel) but DPPA2 overexpression did not induce *Smarca5* expression (Figure 5.29, middle panel). Consistently, *Dppa2* was downregulated in *Smarca5* KO mESCs (Figure 5.29, left panel) but not *vice versa* (Figure 5.31, left panel). These results could indicate either direct regulation of *Dppa2* expression by SMARCA5 or, alternatively, could be explained by a transcriptional feedback loop, since *Dppa2* is also upregulated as part of the 2C-like network and during the major wave of ZGA (Figure 5.26, Appendix I). Analysis of SMARCA5 ChIP-seq data in mESCs could reveal whether it binds directly to the *Dppa2* promoter. However, the SMARCA5 ChIP-seq libraries in mESCs that are publicly available (Local et al. 2018) showed poor genome-wide enrichment over controls (data not shown), consistent with the technical difficulties to ChIP large chromatin remodeling complexes (Iurlaro, M., personal communication).

Finally, given the role of *Dux* in 2C-like and ZGA regulation (Hendrickson et al. 2017; De Iaco et al. 2017; Iaco et al. 2019; Z. Chen & Yi Zhang 2019; Vuoristo et al. 2019) (Figure 1.9) and my previous results showing that both DPPA2 and SMARCA5 strongly induce its expression in mESCs (Figures 5.30 and 5.32), I investigated the link between *Dppa2* and *Dux* using *Dux* KO mESCs (obtained from De Iaco et al. 2017). Since the *Dux* KO mESC line contained an endogenous eGFP reporter, I could not overexpress GFP-fusion proteins. Instead, I transfected an empty vector control or a vector expressing *Dppa2* cDNA for 48 hours and, subsequently, analysed the transcriptome on the whole population by qPCR. *Dppa2* expression levels were not affected in *Dux* KO mESCs, suggesting *Dux* does not regulate *Dppa2* expression (Figure 5.34). As expected, *Dux* expression was undetectable in *Dux* KO mESCs and slightly induced upon DPPA2 overexpression in WT cells (Figure 5.34). However, unlike previous experiments (Figure 5.30), the upregulation of *Dux* by DPPA2 was not statistically significant (Figure 5.34), which could be due to the analysis being done on the whole population of cells rather than on transfected cells selected by FACS sorting. Similar to previous experiments, the pluripotency marker *Pou5f1* (also known as *Oct4*) was unaltered in *Dux* KO mESCs transfected with either an empty vector or DPPA2 (Figure 5.34). As shown in previous studies (Eckersley-Maslin et al. 2019) and in this dissertation (Figures 5.10, 5.15, 5.22-5.24, 5.30), ZGA-like expression was induced downstream of DPPA2 (Figure 5.35). Consistent with the reported role of *Dux* in

ZGA regulation (Hendrickson et al. 2017; De Iaco et al. 2017), *Dux* KO mESCs almost completely abrogated the expression of ZGA genes, and, interestingly, these genes remained expressed at very low levels after these cells were transfected with DPPA2 (Figure 5.35), suggesting DPPA2 acts upstream of DUX. Validating this observation, parallel work to this dissertation showed that the expression levels of ZGA genes in *Dppa2/Dppa4* KO mESCs can be rescued upon DUX overexpression (Eckersley-Maslin et al. 2019).

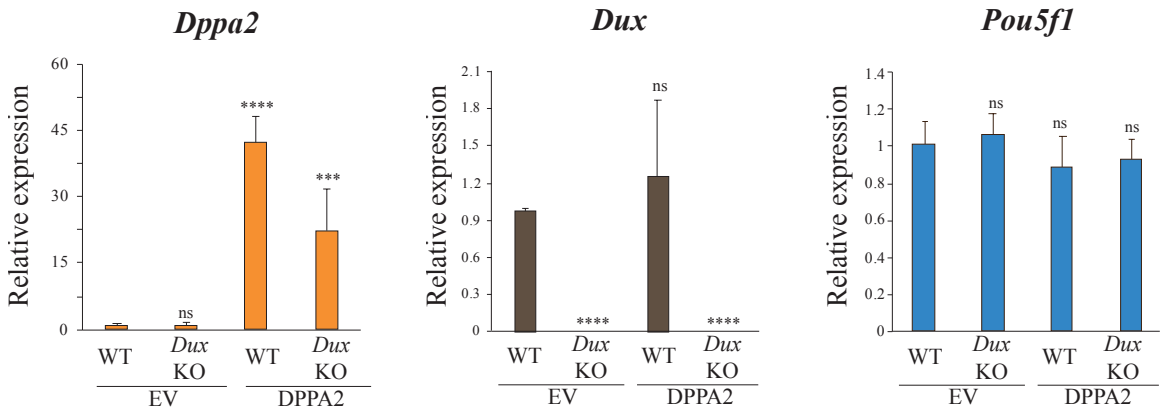



**Figure 5.34– *Dppa2*, *Dux* and *Pou5f1* expression in *Dux* KO mESCs cells transfected with DPPA2, analysed by qPCR**

Analysis of *Dppa2* (left panel, orange), *Dux* (middle panel, brown) and *Pou5f1* (*Oct4*) (right panel, blue) relative mRNA levels by qPCR in wild-type (WT) and *Dux* knock-out (KO) mESCs after 48 hours transient transfection with an empty vector (EV) control or a vector expressing *Dppa2* cDNA. The experiment was performed in three biological replicates and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to WT EV control are reported (homoscedastic two-tailed t-test, ****: p-value <0.0001, ***: p-value <0.001, ns: non-significant).

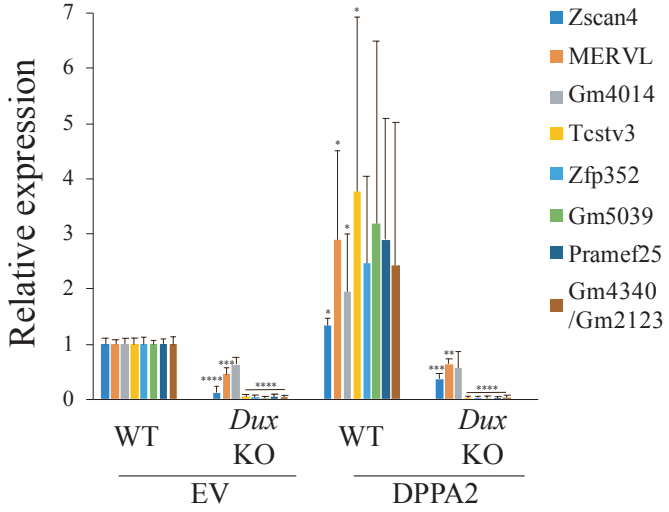



**Figure 5.35– ZGA markers expression in *Dux* KO mESCs transfected with DPPA2, analysed by qPCR**

Analysis of relative RNA levels of a panel of ZGA markers by qPCR in wild-type (WT) and *Dux* knock-out (KO) mESCs after 48 hours transient transfection with an empty vector (EV) control or a vector expressing *Dppa2* cDNA. The experiment was performed in three biological replicates and the qPCR run in two technical replicates. Data is shown as mean plus standard deviation. Statistically significant differences to WT EV control are reported (homoscedastic two-tailed t-test, *: p-value <0.5, **: p-value <0.01, ***: p-value <0.001, ****: p-value <0.0001, absence of starts: non-significant).

Putting together these results, I propose the following model which might explain the interplay between DPPA2, SMARCA5 and DUX during ZGA regulation in mouse embryos (Figure 5.36): at the one-cell or zygote stage, before the major wave of ZGA, SMARCA5 is already present in the pronuclei (Figure 5.27), maintaining a poised open chromatin state of the promoters of ZGA genes and possibly contributing to the low levels of transcription observed during the minor wave of ZGA, including expression of *Dux* transcripts; upon translocation of DPPA2 to the nuclei in the two-cell embryo (Figure 5.27), SMARCA5 facilitates DPPA2 binding to ZGA promoters and, together with DUX, they induce higher levels of transcription during the major wave of ZGA. This model is based on mechanistic studies in mESCs and the expression patterns of these three proteins in the early embryo and, therefore, it remains to be proven *in vivo* (see discussion in section 5.3).

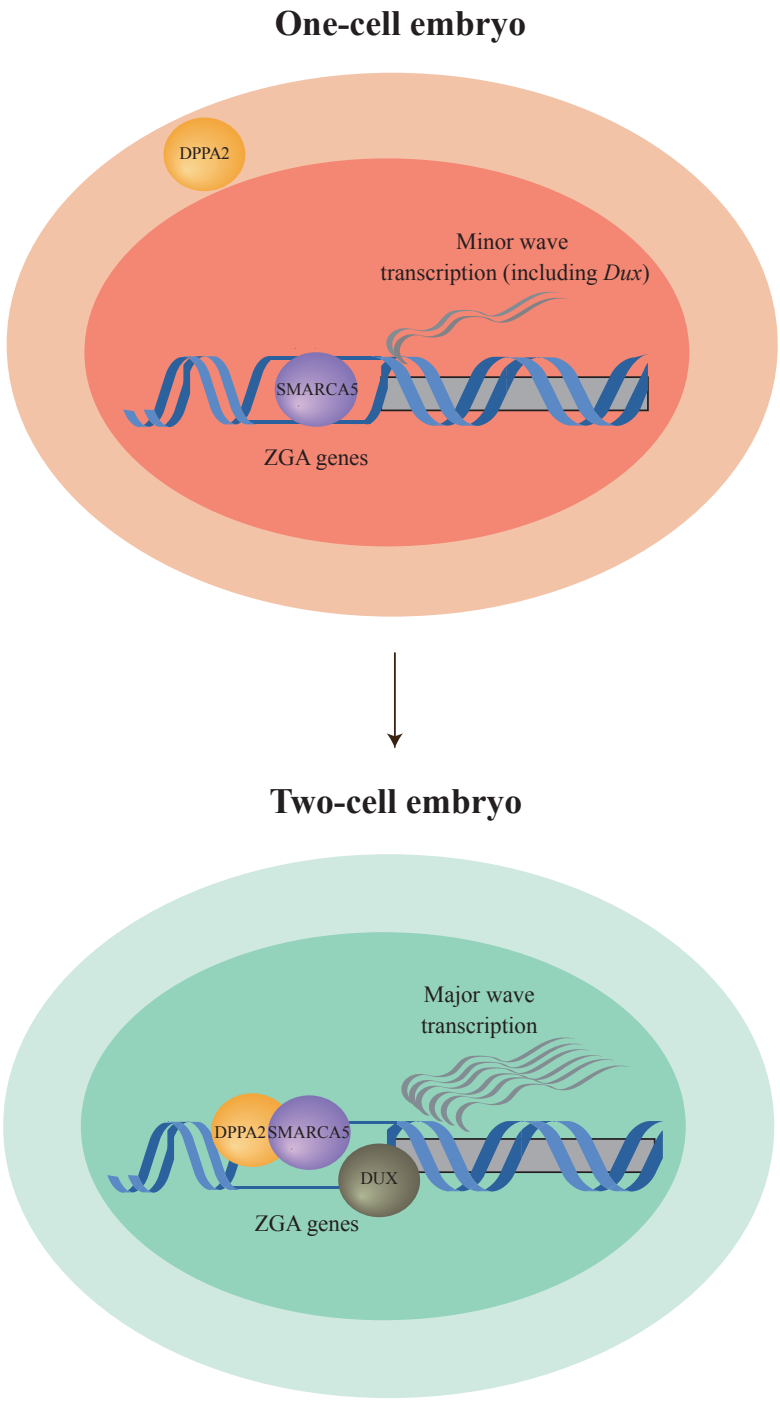



**Figure 5.36– Model of ZGA regulation by DPPA2, SMARCA5 and DUX**

At the one-cell embryo or zygote (top, red), the maternally-inherited DPPA2 (orange) is localised in the cytoplasm (light red area), while in the pronuclei, the chromatin of ZGA gene promoters is bound by the maternal SMARCA5 protein (purple) and transcription occurs and low levels, including expression of *Dux*. The major wave of ZGA is triggered in the two-cell embryo (bottom, green) when DUX is translated (brown) and DPPA2 translocates to the nucleus (dark green) to bind SMARCA5 at ZGA gene promoters. Based on mESC data, DPPA2 and SMARCA5 physically interact (Hernandez et al. 2018), but it is unknown whether DUX physically binds to any of them.

## 5.3 Conclusions and discussion

In this chapter, I analysed the CRISPRa scRNA-seq dataset, generated in chapter 4, where I perturbed 203,894 mESCs to induce the activation of 230 maternally-expressed epigenetic and transcriptional factors, with the aim to identify candidate genes that regulate ZGA-like gene expression *in vitro* (Figure 4.10) and are, therefore, strong candidates for ZGA regulation in mouse embryos. After identifying a prominent ZGA-like signature in the dataset (Figures 5.1-5.3), I collaborated with Oliver Stegle's laboratory to performed integrative dimensionality reduction analysis of the expression of both protein-coding genes and repeat elements (Figures 5.5, 5.7-5.9) to revealed 44 candidate genes that induced a transcriptional response in mESCs reminiscent of the major wave of ZGA (Figure 5.10-5.12). Amongst these factors were *Dppa2*, *Smarca5* and *Patz1*, which I independently validated using complementary experimental approaches (Figures 5.17, 5.19-5.24), also validating my screen design. Finally, I disentangled the interdependencies between *Dppa2*, *Smarca5* and the zygotic transcription factor *Dux* (Figures 5.29-5.33) and put forward a model that might explain how these three factors regulate ZGA in the early mouse embryo (Figure 5.36).

Importantly, while overexpression screens using traditional open reading frame (ORF) libraries have already been coupled with scRNA-seq readout (Parekh et al. 2018), in this dissertation, I present the first screen coupling pooled CRISPRa with single-cell sequencing. As discussed in previous chapters, CRISPRa has key advantages over traditional cDNA overexpression, including target gene upregulation at physiologically-relevant levels (Chavez et al. 2015; Sanson et al. 2018; J. Yang et al. 2019) and the possibility to activate any gene and other genomic features such as repeat elements (Figures 3.5, 3.10, 3.11, 3.17 and 3.21). Consistently, in the validation experiments I performed for *Dppa2*, *Smarca5*, *Patz1* and *Carhsp1*, using both individual CRISPRa and cDNA overexpression, I observed differences in the extent of target gene upregulation induced by either method (Figure 5.19). However, despite these differences in gene dosage and experimental design (Figure 5.17), the transcriptional changes triggered by the screen hits *Dppa2*, *Smarca5* and *Patz1* were considerably similar between methods in the induction of ZGA gene expression (Figures 5.21-5.23), confirming not only that CRISPRa is a robust method to interrogate regulators of ZGA-like transcription but also that these top-

ranking screen hits are potent inductors of ZGA-like transcriptional programmes in mESCs (Figure 5.24). Therefore, *Dppa2*, *Smarca5* and *Patz1* are strong candidates to pursue in *in vivo* experiments.

Screening using single-cell transcriptomics has substantial advantages in terms of scalability and the possibility to disentangle cell-to cell heterogeneity. However, it comes at a cost in terms of sensitivity for detecting transcriptional changes in individual cells (reviewed in Kelsey et al. 2017). To mitigate this reduced sensitivity compared to bulk RNA-sequencing approaches, I considered a ZGA-like transcriptional signature rather than quantification of individual genes for identifying screen hits. This was possible because unsupervised analysis, such as PCA, detected prominent expression of ZGA-like transcripts in the dataset (Figures 5.1-5.3). Consequently, to identify relevant screen hits, in collaboration with Danila Bredikhin and Oliver Stegle, we built a MOFA framework to incorporate both the protein-coding and non-coding transcriptome and define a robust and sensitive signature of ZGA-like transcriptional responses following CRISPRa (Figure 5.5). In this sense, the approach of using a gene signature identified in an unbiased manner rather than conventional differential gene expression analysis might be superior to identify relevant screen hits (Figures 5.10-5.12, 5.15). Firstly, MOFA also considers the expression of repeat elements that are essential during ZGA *in vivo* (see section 1.1.3.3). Secondly, differential gene expression analysis to identify ZGA regulators rely on detecting individual transcripts from a pre-defined list of genes (Figure 5.15) and, therefore, it can introduce biases towards highly expressed genes. In fact, for the three screen hits validated, I showed that most differentially expressed genes estimated based on bulk RNA-sequencing were not differentially expressed in scRNA-seq data and, yet, the transcriptional response captured by both sequencing approaches was remarkably correlated (Figure 5.20), showing the potency of these screen hits in inducing a ZGA-like signature and validating MOFA as a robust method to call screen hits.

To also account for the reduced sensitivity of scRNA-seq, I performed *a priori* power calculations in chapter 3 that allowed me to estimate that at least 367 cells were required per sgRNA to confidently detect a ZGA-like transcriptional response, with the assumption that a positive hit would have similar effects than MERVL and *Zscan4* CRISPRa (Figure 3.24). However, given the sgRNA distribution in the plasmid library before transduction (Figure 4.5),

the number of cells per sgRNA ranged from 9 to 3,300 (Figure 4.16, Appendix C). Within the screen hits identified (Figure 5.10), the number of cells per sgRNA ranged from 50 to 824 (Appendix C). Therefore, it is likely that some of the screen hits with lower cell numbers regulate ZGA-like expression upstream of or via a different transcriptional cascade than MERVL and *Zscan4* and, therefore, less than 367 cells are sufficient to have the power to detect a ZGA-like signature. Nonetheless, before embarking into the molecular characterisation of any of these screen hits, they should be properly validated with alternative methods, such as cDNA overexpression or loss-of-function studies via KO or knock-down. On the same note, it is possible that, within the sgRNAs with lower cell numbers that were not called as hits, there are false negatives that could not be identified due to lack of power.

Even though the utilised MOFA approach could mitigate the described scRNA-seq limitations, detection of target gene activation might have been hampered by the reduced sensitivity in transcript detection in individual cells. While I showed that overall target gene activation and ZGA-like gene expression in the population of cells carrying a specific sgRNA could be detected with less than 10 stringently activated cells (Figures 5.13 and 5.14), I could not detect target gene activation for 20 sgRNA screen hits (Figure 5.14). While I cannot discard the possibility that some of those are false positives, it could also be explained by two other reasons already discussed in chapter 4: 1) technical drop-outs associated with scRNA-seq data, and/or 2) uncoupling of target gene activation and downstream transcriptional responses. These observations further emphasise the need of independent validations for each screen hit. Moreover, these results suggest that the CRISPRa efficiency analysis performed in chapter 4 after categorising effective and not effective sgRNAs based on a threshold of 10 activated cells in the dataset need to be revisited and, overall, more and larger CRISPRa studies coupled with scRNA-seq need to be done to determine robust principles governing CRISPRa efficiency. For instance, it would be interesting to test whether expression of more than one sgRNA targeting the same gene promoter in a single cell would have a synergistic effect on target gene activation and whether that could improve detection by scRNA-seq. In these regards, dual expression vectors could be used to clone two sgRNAs in the same lentiviral backbone under two different U6 promoters (Vidigal & Ventura 2015). However, a Perturb-seq or direct-capture Perturb-seq approach (Dixit et al. 2016; Adamson et al. 2016; Jaitin et al. 2016; Replogle et al. 2018) would

need to be used instead of CROP-seq. This is because the strategy of disrupting the 3' LTR to obtain a polyadenylated transcript containing the sgRNA sequence is limited to a reduced insert size that does not interfere with lentiviral production (Datlinger et al. 2017). Nevertheless, the need to introduce a sgRNA-associated barcode in Perturb-seq-related approaches implies that cloning of dual sgRNA expression vectors needs to be arrayed rather than pooled, likely reducing the throughput of the screen. As an example, a recent studied coupled CRISPRa with scRNA-seq to generate genetic interaction maps in human cells using arrayed cloning of lentiviral Perturb-seq libraries to introduce combinatorial pair-wise combinations of sgRNAs in each cell (Norman et al. 2019).

As discussed earlier, the lower-ranking screen hits were not overall associated with lower total cell numbers, target gene activation or lower enrichment of ZGA genes in the differential gene expression rank (Figures 5.13-5.16, Appendix C). However, the sgRNA obtained as a hit for those lower-ranking gene hits generally showed lower target gene activation than the counterpart sgRNA (Figure 5.14). This suggests that more stringent statistical parameters for calling screen hits that induce ZGA-like expression compared to non-targeting sgRNA controls might be needed to improve the robustness of hit identification and, therefore, the proportion of hits that would be validated both *in vitro* and *in vivo*. Ongoing work aims to explore all these aspects further, including drop-out rates and detection of target gene activation across the screen hit rank.

With the current hit calling approach, out of the 230 maternal candidates screened, 44 were identified as inducers of a ZGA-like response in mESCs (Figures 5.10-5.12, 5.15). Importantly, amongst these hits, MOFA identified factors that had been previously described as ZGA regulators and were, therefore, included in the screen as positive controls, namely *Dppa2* (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Y.-L. Yan et al. 2019), *Yap1* (C. Yu, Ji, Dang, et al. 2016), *Ctcf* (Wan et al. 2008) and *Gata3* (Eckersley-Maslin et al. 2019). This not only validates the screen experimental approach but also the computational framework developed for hit calling. Furthermore, several screen hits, such as *Cbx1*, *Smarca5*, *Phf1*, *Gatad2b*, *Smarca2*, *Hat1*, *Gata3* or *Npm2*, have also been shown to significantly reduce the 2C-like subpopulation in mESCs after siRNA knock-down (Rodriguez-Terrones et al. 2018), consistent with their activator role of ZGA-like expression. Other hits have been implicated in,

but not specifically shown to regulate ZGA transcription. For example, the binding motif of *Arnt* is enriched in distal open chromatin regions of mouse embryos undergoing ZGA (F. Guo et al. 2017). Others, such as *Smarca2*, have been shown to be essential for oogenesis and early embryogenesis in other model organisms (Brizuela et al. 1994). Excitingly, a considerable number of screen hits, including *Patz1*, which I independently validated, have not been previously linked to ZGA. Together, this comparison to previous literature indicates that, despite the potential noise introduced during hit calling, a considerable proportion of the factors identified as inducers of ZGA-like expression in mESCs show great promise to be important regulators of mammalian ZGA. However, other positive controls included for being known regulators of ZGA-like expression in mESCs, such as MERVL, *Zscan4* (as shown in chapter 3) or *Dux* (Hendrickson et al. 2017; De Iaco et al. 2017) were not identified as positive hits. This could be due to two different reasons discussed above: 1) their targeting sgRNAs did not induce sufficient target gene activation, and/or 2) the number of cells expressing their targeting sgRNAs was not sufficient to detect ZGA-like expression with significant statistical power (Appendix C).

Altogether, the study presented in this dissertation opens up exciting avenues for validation and characterisation of the 44 potential ZGA regulators both in mESCs and *in vivo*, testing the functional requirements, interdependencies and redundancies between these new regulators of ZGA-like expression. Of special interest are *Dppa2*, *Smarca5* and *Patz1*, since I have already validated them independently as strong inducers of ZGA gene expression in mESCs (Figures 5.17, 5.22-5.24). These validation experiments suggest that different ZGA regulators might have different mechanisms of action. For instance, while major satellites were overall upregulated in cells expressing the 46 sgRNA hits (Figure 5.12), none of the three validated hits significantly upregulated them compared to controls (Figure 5.23). Additionally, although the genes captured in the MOFA ZGA-like factor were overall upregulated by *Dppa2*, *Smarca5* and *Patz1*, the majority of these genes were induced to different extents by different factors (Figure 5.22). Illustrating these aspects of possible differential ZGA regulation by different screen hits, I showed that *Dppa2* and *Smarca5* present different expression patterns and cellular localisation in early mouse embryos (Figures 5.26 and 5.27). In agreement with the screen results and validation experiments, SMARCA5 localises to sites of active transcription in

zygotes and its zygotic RNAi-mediated knock-down leads to downregulation of some ZGA genes (Torres-Padilla & Zernicka-Goetz 2006). Moreover, *Smarca5* homozygous KO embryos derived from heterozygous crosses arrest during pre-implantation development (Stopka & Skoultchi 2011). Although these studies, together with the results presented here, strongly suggest a role for *Smarca5* during ZGA regulation *in vivo*, depletion of the maternal protein followed by observation of early development and analysis of transcription at the two-cell stage should reveal whether maternal SMARCA5 is truly an essential ZGA regulator. Similar studies are also yet to be done for *Dppa2*, whose ZGA regulatory role has been extensively shown in mESCs (Eckersley-Maslin et al. 2019; De Iaco et al. 2019; Y.-L. Yan et al. 2019) but very little, although promising, evidence supports its role in early embryos. For example, DPPA2-overexpressing mESCs contribute to extraembryonic tissues in chimeric embryos, highlighting the increased developmental potential of these cells (Y.-L. Yan et al. 2019), and injection of a dominant-negative form of DPPA2 into zygotes leads to developmental arrest at the two-cell stage (Hu et al. 2010).

Nevertheless, the accumulation of large pools of proteins during oocyte growth (reviewed in Vastenhouw et al. 2019) makes loss-of-function studies during early development challenging. Different strategies can be used to deplete maternal stores of proteins in early embryos: 1) generation of maternal KO transgenic mouse lines using a Cre-*lox*P system (H. Gu et al. 1993) with a Cre recombinase expressed in oocytes before they enter the growth phase, such as Zp3-Cre (Lewandoski et al. 1997), so that derived embryos are maternal KOs for the protein of interest; 2) siRNA-mediated knock-down *ex vivo* in isolated follicles from young females followed by *in vitro* oocyte growth, maturation and IVF (Eppig & Schroeder 1989; Pfender et al. 2015); and 3) elimination of proteins *ex vivo* using antibody-mediated proteasomal degradation, a method called Trim-away (Clift et al. 2017), in mature oocytes followed by IVF or directly in zygotes. As it will be discussed later in chapter 6, each of these methods have its advantages and disadvantages and need to be carefully consider based on the protein to be studied.

Preliminary data allowed me to propose a model for *Dppa2* and *Smarca5* mechanistic regulation of ZGA in early embryos (Figure 5.36). Using cDNA overexpression approaches in *Dppa2* KO and *Smarca5* KO mESCs, I showed that DPPA2 can induce ZGA-like expression

in the absence of SMARCA5, but the later requires DPPA2 to exerts its ZGA regulatory role (Figures 5.30 and 5.32). Previous observations based on DHS maps of early embryos indicate that transcriptionally inactive genes can have accessible promoters before they become active at later developmental stages (F. Lu et al. 2016). The presence of SMARCA5 in the pronuclei of zygotes (Figure 5.27) at its role as a chromatin remodeler (Barisic et al. 2019) suggest that SMARCA5 might be maintaining a poised open chromatin state of the genes that are to be expressed later on in the major wave of ZGA. While there could be redundancy in this function between different chromatin remodellers such as *Smarca2* or *Arid1a*, both identified as screen hits, the other ATPase subunit of the ISWI complex, *Smarca1* (also known as SNF2L), could not play a role during ZGA regulation since it is not expressed in oocytes or early mouse embryos (Pfeiffer et al. 2011; Pfeiffer et al. 2015; B. Wang et al. 2016). This potential redundancy is supported by the observation that *Smarca5* KO mESCs downregulated, but did not completely abolish, expression of ZGA genes (Figure 5.30). Nevertheless, the hypothesis that SMARCA5 regulates the chromatin state of ZGA genes in early embryos needs to be proven with ATAC-seq or DNase-seq approaches upon depletion of the maternal protein. Additionally, low-input SMARCA5 ChIP-seq or single-cell CUT&Tag (Kaya-Okur et al. 2019) in early embryos should reveal whether SMARCA5 binds genes in zygotes that become expressed later on at the two-cell stage; however, as discussed earlier, this might be challenging due to the difficulties to ChIP large chromatin complexes. Moreover, SMARCA5 has been shown to regulate TAD formation in mESCs (Barisic et al. 2019); single-cell Hi-C analysis on embryos depleted of the maternal protein would unravel whether this mechanism also applies to the MZT.

Importantly, low levels of transcription are detected during the minor wave of ZGA in the paternal pronuclei of one-cell embryos (Ram & Schultz 1993; Bouniol et al. 1995; Majumder & DePamphilis 1995; Nothias et al. 1996; Aoki et al. 1997; F. Zeng et al. 2004; Xue et al. 2013; Deng et al. 2014; Abe et al. 2015; S.-J. Park et al. 2015). Whether SMARCA5 contributes to the minor wave of ZGA and, if it does, what are the factors that protect the maternal pronuclei, where SMARCA5 is also expressed (Figure 5.27), from being transcribed are questions that remained to be answered. Strikingly, SMARCA5 upregulates *Dux* in mESCs (Figure 5.32), which is transcribed during the minor wave of ZGA (De Iaco et al. 2017;

Hendrickson et al. 2017), suggesting this chromatin remodeler might indeed have a function during zygotic transcription.

In mESCs, SMARCA5 needs DPPA2 to induce ZGA-like expression (Figures 5.30 and 5.32). In mouse embryos, the trigger of the major wave of ZGA could be the translocation of DPPA2 from the cytoplasm to the nucleus at the two-cell stage and expression DUX (Figures 5.27 and 5.36). How this change in DPPA2 localisation is regulated remains an open question. In mESCs, a post-translational modification of DPPA2, sumoylation mediated by PIAS4, regulates DPPA2 heterodimerisation to DPPA4, subsequent transcription of *Dux* and downstream activation of a ZGA-like programme (Y.-L. Yan et al. 2019). It will be exciting to uncover whether such mechanisms apply in the early embryo and how they relate to DPPA2 localisation, specially in relation to SMARCA5. This could be addressed with loss-of-function studies of *Pias4*, which is also expressed as a maternal protein (Pfeiffer et al. 2011; Pfeiffer et al. 2015; B. Wang et al. 2016), followed by immunofluorescence analysis of early embryos. Notably, *Dppa4* has been shown to regulate ZGA together with *Dppa2* in mESCs (Eckersley-Maslin et al. 2019; De Iaco et al. 2019). Unfortunately, I could not verify *Dppa4* in my screen, possibly due to their sgRNAs mediating insufficient target gene activation (Appendix C). However, these future functional studies proposed here should also consider *Dppa4* expression and localisation in the early embryo for a comprehensive picture of ZGA regulation.

Further evidence that supports the proposed model for DPPA2 and SMARCA5 regulation of ZGA (Figure 5.36) comes from analysis of LINE-1 expression after DPPA2 and SMARCA5 overexpression. LINE-1 represses ZGA expression after the two-cell stage via chromatin remodeling (Jachowicz et al. 2017; Percharde et al. 2018). Consistent with a role for DPPA2 in activating transcription during the major wave of ZGA and SMARCA5 merely facilitating it (Figure 5.36), LINE-1 is upregulated by DPPA2 but not by SMARCA5 (Figure 5.23), suggesting DPPA2 contributes to LINE-1 activation after ZGA is triggered to help repressing it later on. DPPA2-mediated regulation of LINE-1 has also been previously shown (Eckersley-Maslin et al. 2019; De Iaco et al. 2019) and suggests a negative feedback loop that could explain why DPPA2, despite being expressed in all mESCs, only activates ZGA-like expression in a subset of them (Eckersley-Maslin et al. 2019; De Iaco et al. 2019). An alternative explanation to this observation is that DPPA2 needs SMARCA5 to set a specific chromatin state in mESCs

to induce ZGA-like expression. However, SMARCA5 is also constitutively expressed in mESCs (Figure 5.18) and, therefore, there might be additional mechanisms that regulate the entry to the 2C-like state in mESCs. In this sense, co-overexpression of DPPA2 and SMARCA5 in mESCs followed by transcriptional analysis of ZGA genes should be performed to validate whether SMARCA5 and DPPA2 have synergistic functions in ZGA regulation. Notably, immunoprecipitation experiments showed that DPPA2 and SMARCA5 physically interact in mESCs (Hernandez et al. 2018). In the embryo, co-localisation analysis showed that these two regulators co-localise in the nuclei of two-cell embryos (Figures 5.27 and 5.28). Although these results should be validated by super-resolution imaging, they could indicate that SMARCA5 helps recruit DPPA2 to ZGA gene promoters in two-cell embryos. To prove this and other hypothesis proposed here, similar mechanistic experiments to those performed in this chapter in mESCs should now be done in early embryos using maternal KOs for *Dppa2*, *Smarca5* and *Dux* to understand how they regulate each other and ZGA.

# Chapter 6 Conclusions and outlook

In this dissertation, I combined state-of-the-art technologies to develop a novel genetic screening strategy in mammalian cells. Since the discovery of CRISPR as a tool for genetic engineering in 2012 (Jinek et al. 2012; Gasiunas et al. 2012), this potent technology has been widely used for genetic screening applied to a vast range of research questions and across multiple model organisms (reviewed in Doench 2018). Nevertheless, it was not until 2016 that complex read-outs were implemented in pooled CRISPR screens by the combination with single-cell sequencing (Dixit et al. 2016; Adamson et al. 2016; Jaitin et al. 2016; Datlinger et al. 2017), allowing comprehensive interrogation of gene function and regulation at a cellular level in an unbiased manner. This powerful screening technology, implemented initially for loss-of-function perturbations, has been successfully applied in independent studies during the last few years (S. Xie et al. 2017; Replogle et al. 2018; Rubin et al. 2019; Gasperini et al. 2019). However, the proposal and development of gain-of-function perturbations mediated by CRISPR (CRISPRa) and assessed by single-cell sequencing had not been addressed until now. At the same time that the work presented in this dissertation was submitted for publication, a study was published describing the generation of genetic interaction maps in human cells by combining combinatorial CRISPRa with scRNA-seq (Norman et al. 2019). Nonetheless, the concept of coupling pooled CRISPRa with scRNA-seq for screening purposes is presented as novel in this dissertation. My novel screening method provides a powerful way to systematically interrogate a large number of genes for their effects on specific transcriptional programmes, such as ZGA, and I anticipate that it will be widely adaptable and used in the future to answer many other research questions in different biological contexts.

Given the transcriptionally inactive state prior to ZGA in the embryo and the extensive epigenetic reprogramming occurring at the same time, the identification of maternally-deposited epigenetic and transcriptional factors that induce ZGA transcription is crucial to understand this critical event in early development (reviewed in Eckersley-Maslin, Alda-

Catalinas & Reik 2018). Due to the limitations of high-throughput screening in pre-implantation mouse embryos, I turned to mESCs reasoning that activating ZGA regulators would induce ZGA-like transcription. In fact, this concept had been previously proven in a candidate-based screen using cDNA overexpression followed by qPCR and flow cytometry as read-outs (Eckersley-Maslin et al. 2019). Pooled CRISPRa combined with single-cell transcriptomics presented itself as an ideal technology to interrogate multiple factors for their role as ZGA regulators. I reasoned that the use of pooled CRISPRa would considerably increase the throughput of the screen and, consequently, the number of factors that could be interrogated, and scRNA-seq would allow comprehensive interrogation of the transcriptional changes triggered by the screen candidates.

Therefore, I set out to optimise CRISPRa in mESCs using MERVL and the *Zscan4* cluster as targets and showed that their activation in mESCs induced a ZGA-like response that could be captured by scRNA-seq (Figure 3.24 and 3.25). This initial pilot test allowed me to design a screen for 230 maternal epigenetic and transcriptional factors (Figure 4.10). I overcame the limitation of capturing sgRNAs in scRNA-seq libraries by adapting the CROP-seq strategy (Datlinger et al. 2017) to CRISPRa SAM (Figure 3.27), one of the most potent CRISPRa technologies developed to date (Konermann et al. 2015; Chavez et al. 2016). After transducing a lentiviral library containing two sgRNAs per candidate gene into mESCs expressing CRISPRa SAM machinery (SAM22 mESCs) and generating a scRNA-seq dataset, I first aimed to assess the principles governing CRISPRa efficiency. The most striking observation I made was that, within the 180 bp targeting window upstream of the gene TSS, previously defined to be optimal for CRISPRa SAM (Konermann et al. 2015), effective sgRNAs mapped further away from the gene TSS than the ineffective ones (Figure 4.23). Next, in collaboration with Oliver Stegle's laboratory (EMBL, Heidelberg), we built a computational framework based on integrative dimensionality reduction analysis that allowed the identification of 46 sgRNAs targeting 44 unique candidate genes whose CRISPRa induced a ZGA-like transcriptional response in mESCs (Figure 5.10). After validation of three of the top-ranking hits, *Dppa2*, *Smarca5* and *Patz1* (Figures 5.22-5.24), I focused on the mechanistic regulation of *Dppa2* and *Smarca5* and proposed a model explaining their potential mechanism of action during ZGA regulation in mouse embryos (Figure 5.36).

By identifying novel epigenetic and transcriptional regulators of ZGA, this dissertation not only opens up multiple new lines of research but also constitutes a step forward towards the understanding of the relationship between chromatin remodeling and transcription in early embryos. However, detailed functional studies should be done in the future to understand the mechanistic regulation of these factors, both in relation to each other but also in terms of the specific transcriptional networks that they activate. In this sense, many questions remain unanswered: is there redundancy in the regulation of ZGA? how many factors are necessary to trigger ZGA? do these factors regulate the same genes and how many act via MERVL and/or *Zscan4* activation? is epigenetic reprogramming necessary for ZGA to occur or is ZGA a consequence of this epigenetic remodeling? what is the function, if any, of the genes transcribed during ZGA? Understanding the mechanism of action of these newly identified regulators should provide some answers to these questions.

However, before embarking into deeper mechanistic studies, these candidates should be first individually validated using alternative approaches. The most straight forward *in vitro* validation approach is probably the one presented in this dissertation, consisting of individual CRISPRa and cDNA overexpression of the candidates followed by RNA-sequencing or, simply, analysis of a panel of ZGA genes by qPCR. Nonetheless, it is important to consider that *in vitro* screening has the advantage of interrogating multiple genes and generating rich-information results, but does not necessarily guarantee that the identified hits play the same role in the embryo and, therefore, *in vivo* studies are necessary. This is the example of *Dux*, which was identified as a key regulator of ZGA-like expression in mouse and human ESCs (Hendrickson et al. 2017; De Iaco et al. 2017) but *in vivo* studies revealed that it is not essential for development (Iaco et al. 2019; Z. Chen & Yi Zhang 2019). Notably, rather than indicating that *Dux* does not function during ZGA, these observations could also point towards redundancy in ZGA regulation, where multiple factors are needed for it to occur properly but, in the absence of one of them, others can compensate for its function.

As discussed in chapter 5, the accumulation of large pools of proteins during oocyte growth makes loss-of-function studies in early embryos challenging. These can be addressed using three different strategies that can be complementary to each other: 1) transgenic mouse lines of conditional maternal KOs generated using a Cre-*lox*P system with an oocyte-specific Cre

(Lewandoski et al. 1997); 2) siRNA-mediated knock-down *ex vivo* in isolated follicles followed by *in vitro* oocyte growth, maturation and IVF (Eppig & Schroeder 1989; Pfender et al. 2015); and 3) protein depletion *ex vivo* using Trim-away (Clift et al. 2017) for antibody-mediated proteasomal degradation in mature oocytes followed by IVF or, directly, in zygotes. All these approaches require detailed knowledge of when the protein is expressed during folliculogenesis and oogenesis to target it at the right time. Transgenic deletion of maternal proteins to study early development is only feasible if the protein of interest is not involved in oogenesis. Moreover, while this approach is time- and cost-consuming, the other two approaches require complex micromanipulation techniques. siRNA-mediated knock-down in follicles requires effective siRNAs which should be tested in cells before undertaking complex *in vivo* studies. Similarly, Trim-away requires a specific antibody, which should also be tested beforehand *in vitro*. Altogether, one needs to balance the pros and cons of each of these strategies and decide on the best way to study maternal proteins in early development. Currently ongoing work together with Oana Kubinyecz and Melanie Eckersley-Maslin in the Reik laboratory aims to validate and understand *in vivo* the role of *Dppa2* and *Smarca5* during early development, specifically during ZGA regulation. Conditional KO mouse lines for these genes have been generated and are currently being crossed to Zp3-Cre transgenic lines to create maternal KO lines. If this approach turned out to be unsuitable because either of these proteins are involved in oocyte development, a Trim-away approach is currently being optimised using already validated antibodies. These studies should reveal whether *Dppa2* and/or *Smarca5* are essential ZGA regulators, and mechanistic studies downstream of protein depletion in oocytes should shed light into their mechanism of action, including potential interdependencies between them and regulation of the transcription factor *Dux*.

Alternatively, instead of testing individual screen hits, one could try to narrow down the list of factors to study by performing a downstream combinatorial screen. This approach was not feasible to interrogate the 230 factors tested in this dissertation, since many more cells would have had to be sequenced, making the study cost-ineffective. However, with the list of candidates narrowed down to 44, it is possible to design combinatorial approaches as a way to study functional relationships between them. For that, multiple sgRNAs targeting different genes can be introduced into SAM22 mESCs using a high MOI transduction of the lentiviral

sgRNA library. Then, transcriptional analysis by scRNA-seq would reveal which CRISPRa combinations render a synergistic, antagonistic or neutral effect on ZGA-like expression.

Even though I focused on epigenetic and transcriptional regulators, it is possible that proteins with other functions regulate ZGA directly or indirectly. Similar screens to the one described in this dissertation could be performed to interrogate other functionalities of ZGA regulation. Additionally, the sgRNA design could be improved for these future screens using the observations reported here and in other studies (chapter 4, Radzisheuskaya et al. 2016; Horlbeck et al. 2016; Sanson et al. 2018).

Furthermore, in the future, it will also be interesting to see how these findings translate to human ZGA: are the regulators identified here expressed during early human development? can a human ZGA-like signature be induced in human ESCs and detected by scRNA-seq to perform similar screens? DUX4 overexpression in human ESCs suggests that similar screens might be possible (De Iaco et al. 2017; Hendrickson et al. 2017), however, other potential factors should be tested individually before performing large-scale screens.

Moreover, pooled CRISPRa screening could be combined with other single-cell read-outs. As a proof-of-concept, pooled CRISPRi has been combined with single-cell-ATAC-seq (Rubin et al. 2019). Although this technology is currently not high-throughput, the recent development of single-cell-ATAC-seq in droplet-based platforms (Satpathy et al. 2019) promises exciting avenues for the development of high-throughput CRISPR screening with an epigenetic read-out. The application of such strategy to interrogate the epigenetic functions of ZGA regulators *in vitro* would be extremely interesting, given the closed link between transcription and epigenetic remodeling during early development.

Lastly, this study also opens up new avenues for the development of novel computational approaches, as shown here with the application of MOFA. For example, ongoing work aims to gain detailed insights into the transcriptional networks triggered by each screen hit and the cross-talk between them. Additionally, the main drawback I encountered in this screen was the lack of detection of target gene activation for a considerable number of factors (Appendix C). While this could be due to lack of transcript detection explained by drop-outs associated with scRNA-seq data, moving forward, it will be important to set ground rules to define effective

and ineffective sgRNAs. More CRISPRa scRNA-seq studies, which I anticipate will become available in the near future, should be systematically analysed to set these rules and comprehensively assess genomic and sgRNA features that influence CRISPRa efficiency. Importantly, these efficiency rules should be tested across different CRISPRa technologies and mammalian cell lines. This will help towards the development of improved genome-wide and targeted CRISPRa libraries. The dataset presented in this study constitutes a starting point to systematically analyse drop-out rates in relation to CRISPRa SAM efficiency, define the minimum number of cells required to detect target gene activation and how that relates to basal gene expression levels of the targeted gene and other epigenetic features. Even though such analysis were done preliminarily in chapter 4 of this dissertation, the sgRNA classification based on the threshold of 10 activated cells should be revisited in light of drop-out analysis. Furthermore, machine learning algorithms could be implemented using this dataset considering sgRNA features, basal transcriptional and epigenetic state of the target genes, detection of target gene activation, drop-out rates and cell cycle state of each individual cell, amogst other features, to provide a better understanding of CRISPRa efficiency.

In summary, I conclude that my CRISPRa single-cell transcriptomics screen has unraveled novel positive regulators of ZGA. Furthermore, I showed that this novel screening method provides a powerful way to systematically interrogate a large number of genes for their effects on specific transcriptional programmes. In the future, *in vitro* and *in vivo* experiments to test the functional requirements, interdependencies and redundancies between these new regulators, additional complementary screens for ZGA regulation, adaptation of pooled CRISPRa screening to other single-cell read-outs, and development of novel computational approaches to analyse these rich-information and complex datasets show great promise to continue expanding our knowledge of mammalian gene function, with special emphasis on the critical events that occur in early development.

# Chapter 7 Bibliography


Abe, K.-I. et al., 2018. Minor zygotic gene activation is essential for mouse preimplantation development. *Proceedings of the National Academy of Sciences*, 115(29), E6780-R6788

Abe, K.I. et al., 2015. The first murine zygotic transcription is promiscuous and uncoupled from splicing and 3' processing. *The EMBO journal*, 34(11), pp.1523–1537.

Adamson, B. et al., 2016. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7), pp.1867–1882.e21.

Adamson, B. et al., 2018. Approaches to maximize sgRNA-barcode coupling in Perturb-seq screens. *bioRxiv*, p.298349.

Adli, M., 2018. The CRISPR tool kit for genome editing and beyond. *Nature Communications*, 9(1), p.1911.

Akiyama, T. et al., 2011. Dynamic replacement of histone H3 variants reprograms epigenetic marks in early mouse embryos. *PLOS Genetics*, 7(10), p.e1002279.

Akiyama, T. et al., 2015. Transient bursts of Zscan4 expression are accompanied by the rapid derepression of heterochromatin in mouse embryonic stem cells. *DNA Research*, 22(5), pp. 313–318.

Alizadeh, Z., Kageyama, S.-I. & Aoki, F., 2005. Degradation of maternal mRNA in mouse embryos: selective degradation of specific mRNAs after fertilization. *Mol Reprod Dev*, 72(3), pp.281–290.

Allis, C.D. & Jenuwein, T., 2016. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, 17(8), pp.487–500.

Amano, T. et al., 2013. Zscan4 restores the developmental potency of embryonic stem cells. *Nature Communications*, 4:1996

Ambriz, X., de Lanerolle, P. & Ambrosio, J.R., 2018. The Mechanobiology of the Actin Cytoskeleton in Stem Cells during Differentiation and Interaction with Biomaterials. *Stem Cells International*, 2018, pp.2891957–2891957.

Amleh, A. & Dean, J., 2002. Mouse genetics provides insight into folliculogenesis, fertilization and early embryonic development. *Human Reproduction Update*, 8(5), pp.395–403.

Amouroux, R. et al., 2016. De novo DNA methylation drives 5hmC accumulation in mouse zygotes. *Nature cell biology*, 18(2), pp.225–233.

Ancelin, K. et al., 2016. Maternal LSD1/KDM1A is an essential regulator of chromatin and transcription landscapes during zygotic genome activation. *eLife*, 5, e08851

Andreu-Vieyra, C.V. et al., 2010. MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. *PLoS Biol*, 8(8).

Angermueller, C. et al., 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3), pp.229–232.

Aoki, F., Worrad, D.M. & Schultz, R.M., 1997. Regulation of transcriptional activity during the first and second cell cycles in the preimplantation mouse embryo. *Developmental Biology*, 181(2), pp.296–307.

Aoshima, K. et al., 2015. Paternal H3K4 methylation is required for minor zygotic gene activation and early mouse embryonic development. *EMBO reports*, 16(7), pp.803–812.

Arakawa, T. et al., 2015. Stella controls chromocenter formation through regulation of Daxx expression in 2-cell embryos. *Biochemical and Biophysical Research Communications*, 466(1), pp.60–65.

Argelaguet, R. et al., 2018. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, 14(6), p.e8124.

Argelaguet, R. et al., 2019. Single cell multi-omics profiling reveals a hierarchical epigenetic landscape during mammalian germ layer specification. *bioRxiv*, p.519207.

Arnoult, C. et al., 1996. Activation of mouse sperm T-type Ca2+ channels by adhesion to the egg zona pellucida. *Proceedings of the National Academy of Sciences*, 93(23), pp.13004–13009.

Au Yeung, W.K. et al., 2019. Histone H3K9 Methyltransferase G9a in Oocytes Is Essential for Preimplantation Development but Dispensable for CG Methylation Protection. *Cell Reports*, 27(1), pp.282–293.e4.

Azuara, V. et al., 2006. Chromatin signatures of pluripotent cell lines. *Nature cell biology*, 8(5), pp.532–538.

Bailey, T.L. et al., 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue), pp.W202–8.

Baker, C.L. & Pera, M.F., 2018. Capturing Totipotent Stem Cells. *Cell Stem Cell*, 22(1), pp.25–34.

Bannister, A.J. et al., 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, 410(6824), pp.120–124.

Barau, J. et al., 2016. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science*, 354(6314), pp.909–912.

Barisic, D. et al., 2019. Mammalian ISWI and SWI/SNF selectively mediate binding of distinct transcription factors. *Nature*, 569(7754), pp.136-140

Barrangou, R. et al., 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819), pp.1709–1712.

Bazzini, A.A. et al., 2016. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *The EMBO journal*, 35(19), pp.2087–2103.

Behjati, S. et al., 2018. Mapping human development at single-cell resolution. *Development*, 145(3).

Beraldi, R. et al., 2006. Expression of LINE-1 retroposons is essential for murine preimplantation development. *Mol Reprod Dev*, 73(3), pp.279–287.

Bernstein, B.E. et al., 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2), pp.315–326.

Berrens, R.V. et al., 2017. An endosiRNA-Based Repression Mechanism Counteracts Transposon Activation during Global DNA Demethylation in Embryonic Stem Cells. *Cell Stem Cell*, 21(5), pp.694–703.e7.

Bestor, T.H., 2000. The DNA methyltransferases of mammals. *Human Molecular Genetics*, 9(16), pp.2395–2402.

Biase, F.H., Cao, X. & Zhong, S., 2014. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Research*, 24(11), pp.1787–1796.

Bilodeau, S. et al., 2009. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes & Development*, 23(21), pp.2484–2489.

Bird, A., 2002. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1), pp.6–21.

Bird, A. et al., 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1), pp.91–99.

Birnbaum, K.D., 2018. Power in Numbers: Single-Cell RNA-Seq Strategies to Dissect Complex Tissues. *Annu Rev Genet*, 52, pp.203–221.

Black, J.B. et al., 2016. Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells. *Cell Stem Cell*, 19(3), pp.406–414.

Blakeley, P. et al., 2015. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development*, 142(18), pp.3151–3165.

Bleil, J.D. & Wassarman, P.M., 1980. Mammalian sperm-egg interaction: identification of a glycoprotein in mouse egg zonae pellucidae possessing receptor activity for sperm. *Cell*, 20(3), pp.873–882.

Boiani, M. et al., 2019. Totipotency continuity from zygote to early blastomeres: a model under revision. *Reproduction*, 158(2), pp.R49–R65.

Bolotin, A. et al., 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology (Reading, England)*, 151(Pt 8), pp.2551–2561.

Bonev, B. & Cavalli, G., 2016. Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11), pp.661–678.

Boraas, L.C., Pineda, E.T. & Ahsan, T., 2018. Actin and myosin II modulate differentiation of pluripotent stem cells. *PLoS ONE*, 13(4), p.e0195588.

Boroviak, T. et al., 2018. Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development*, 145(21).

Borsos, M. & Torres-Padilla, M.-E., 2016. Building up the nucleus: nuclear organization in the establishment of totipotency and pluripotency during mammalian development. *Genes & Development*, 30(6), pp.611–621.

Borsos, M. et al., 2019. Genome-lamina interactions are established de novo in the early mouse embryo. *Nature*, 569(7758), pp.729-733

Bortvin, A. et al., 2004. Dppa3 / Pgc7 / stella is a maternal factor and is not required for germ cell specification in mice. *BMC Developmental Biology*, 4, p.2.

Bortvin, A. et al., 2003. Incomplete reactivation of Oct4-related genes in mouse embryos cloned from somatic nuclei. *Development*, 130(8), pp.1673–1680.

Bose, S. et al., 2015. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome biology*, 16, p.120.

Bošković, A. et al., 2014. Higher chromatin mobility supports totipotency and precedes pluripotency in vivo. *Genes & Development*, 28(10), pp.1042–1047.

Bouniol, C., Nguyen, E. & Debey, P., 1995. Endogenous transcription occurs at the 1-cell stage in the mouse embryo. *Experimental cell research*, 218(1), pp.57–62.

Bourc'his, D. et al., 2001. Dnmt3L and the establishment of maternal genomic imprints. *Science*, 294(5551), pp.2536–2539.

Bourque, G. et al., 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research*, 18(11), pp.1752–1762.

Boyer, L.A. et al., 2000. Functional delineation of three groups of the ATP-dependent family of chromatin remodeling enzymes. *Journal of Biological Chemistry*, 275(25), pp.18864–18870.

Braude, P., Bolton, V. & Moore, S., 1988. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. Nature, 332(6163), pp.459–461.

Braun, C.J. et al., 2016. Versatile in vivo regulation of tumor phenotypes by dCas9-mediated transcriptional perturbation. *Proc Natl Acad Sci USA*, 113(27), p.E3892.

Brennecke, P. et al., 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11), pp.1093–1095.

Brici, D. et al., 2017. The histone 3 lysine 4 methyltransferase Setd1b is a maternal effect gene required for the oogenic gene expression program. *Development*.

Brizuela, B.J. et al., 1994. Genetic analysis of the brahma gene of Drosophila melanogaster and polytene chromosome subdivisions 72AB. *Genetics*, 137(3), pp.803–813.

Brouns, S.J.J. et al., 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, 321(5891), pp.960–964.

Buenrostro, J.D. et al., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12), pp.1213–1218.

Buisson, M. et al., 1989. The Epstein-Barr virus (EBV) early protein EB2 is a posttranscriptional activator expressed under the control of EBV transcription factors EB1 and R. *Journal of virology*, 63(12), pp.5276–5284.

Bultman, S.J. et al., 2006. Maternal BRG1 regulates zygotic genome activation in the mouse. *Genes & Development*, 20(13), pp.1744–1754.

Burns, K.H. et al., 2003. Roles of NPM2 in chromatin and nucleolar organization in oocytes and embryos. *Science*, 300(5619), pp.633–636.

Burton, A. & Torres-Padilla, M.-E., 2010. Epigenetic reprogramming and development: a unique heterochromatin organization in the preimplantation mouse embryo. *Briefings in functional genomics*, 9(5-6), pp.444–454.

Cao, J. et al., 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352), pp.661–667.

Cao, J. et al., 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745), pp.496–502.

Casanova, M. et al., 2013. Heterochromatin Reorganization during Early Mouse Development Requires a Single-Stranded Noncoding Transcript. *Cell Reports*, 4(6), pp.1156–1167.

Casser, E. et al., 2018. Retrospective analysis: reproducibility of interblastomere differences of mRNA expression in 2-cell stage mouse embryos is remarkably poor due to combinatorial mechanisms of blastomere diversification. *Molecular Human Reproduction*, 24(7), pp.388–400.

Casser, E. et al., 2017. Totipotency segregates between the sister blastomeres of two-cell stage mouse embryos. *Science Reports*, 7(1), p.8299.

Chakrabarti, A.M. et al., 2019. Target-Specific Precision of CRISPR-Mediated Genome Editing. *Molecular Cell*, 73(4), pp.699–713.e6.

Chakraborty, S. et al., 2014. A CRISPR/Cas9-Based System for Reprogramming Cell Lineage Specification. *Stem Cell Reports*, 3(6), pp.940–947.

Chang, H. et al., 2018. Terminal Uridylyltransferases Execute Programmed Clearance of Maternal Transcriptome in Vertebrate Embryos. *Molecular Cell*, 70(1), pp.72–82.e7.

Chapman, A.R. et al., 2015. Single cell transcriptome amplification with MALBAC. *PLoS ONE*, 10(3), p.e0120889.

Chari, R. et al., 2015. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nature Methods*, 12(9), pp.823–826.

Chavez, A. et al., 2016. Comparison of Cas9 activators in multiple species. *Nature Methods*, 13(7), pp.563–567.

Chavez, A. et al., 2015. Highly efficient Cas9-mediated transcriptional programming. *Nature Methods*, 12(4), pp.326–328.

Chen, B., Guan, J. & Huang, B., 2016. Imaging Specific Genomic DNA in Living Cells. *Annual review of biophysics*, 45, pp.1–23.

Chen, J. et al., 2011. Genome-wide analysis of translation reveals a critical role for deleted in azoospermia-like (Dazl) at the oocyte-to-zygote transition. *Genes & Development*, 25(7), pp.755–766.

Chen, K.H. et al., 2015. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233), pp.aaa6090–aaa6090.

Chen, Z. & Zhang, Yi, 2019. Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nature Genetics*.

Cheng, A.W. et al., 2013. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Research*, 23(10), pp.1163–1171.

Choi, Y.J. et al., 2017. Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells. *Science*, 355(6325).

Choudhury, S.R. et al., 2016. CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. *Oncotarget*, 7(29), pp.46545–46556.

Christians, E. et al., 2000. Maternal effect of Hsf1 on reproductive success. *Nature.*, 407(6805), pp.693–694.

Clark, S.J. et al., 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9(1), p.781.

Clegg, K.B. & Piko, L., 1983. Poly(A) length, cytoplasmic adenylation and synthesis of poly(A)+ RNA in early mouse embryos. *Developmental Biology*, 95(2), pp.331–341.

Clift, D. et al., 2017. A Method for the Acute and Rapid Degradation of Endogenous Proteins. *Cell*, 171(7), pp.1692–1706.e18.

Cong, L. et al., 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121), pp.819–823.

Crawford, G.E. et al., 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, 16(1), pp.123–131.

Dahl, J.A. et al., 2016. Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature*, 537(7621), pp.548–552.

Dan, J. et al., 2015. Roles for Histone Acetylation in Regulation of Telomere Elongation and Two-cell State in Mouse ES Cells. *Journal of Cellular Physiology*, 230(10), pp.2337–2344.

Dan, J. et al., 2013. Roles for Tbx3 in regulation of two-cell state and telomere elongation in mouse ES cells. *Science Reports*, 3, pp.1–9.

Dan, J. et al., 2017. Zscan4 Inhibits Maintenance DNA Methylation to Facilitate Telomere Elongation in Mouse Embryonic Stem Cells. *Cell Reports*, 20(8), pp.1936–1949.

Dang-Nguyen, T.Q. & Torres-Padilla, M.-E., 2015. How cells build totipotency and pluripotency: nuclear, chromatin and transcriptional architecture. *Current Opinion in Cell Biology*, 34, pp.9–15.

Datlinger, P. et al., 2017. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3), pp.297–301.

Davis, W., Jr, De Sousa, P.A. & Schultz, R.M., 1996. Transient Expression of Translation Initiation Factor eIF-4C during the 2-Cell Stage of the Preimplantation Mouse Embryo: Identification by mRNA Differential Display and the Role of DNA Replication in Zygotic Gene Activation. *Developmental Biology*, 174(2), pp.190–201.

Dayalan Naidu, S. & Dinkova-Kostova, A.T., 2017. Regulation of the mammalian heat shock factor 1. *The FEBS journal*, 284(11), pp.1606–1627.

De Iaco, A. et al., 2019. DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. *EMBO Rep*, p.e47382.

De Iaco, A. et al., 2017. DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nature Genetics*, 49(6), pp.941-945

De Leon, V., Johnson, A. & Bachvarova, R., 1983. Half-lives and relative amounts of stored and polysomal ribosomes and poly(A) + RNA in mouse oocytes. *Developmental Biology*, 98(2), pp.400–408.

De Los Angeles, A. et al., 2015. Hallmarks of pluripotency. *Nature*, 525(7570), pp.469–478.

de Vries, W.N. et al., 2004. Maternal beta-catenin and E-cadherin in mouse development. *Development*, 131(18), pp.4435–4445.

Dekker, J., Marti-Renom, M.A. & Mirny, L.A., 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6), pp.390–403.

Deltcheva, E. et al., 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, 471(7340), pp.602–607.

Deng, Q. et al., 2014. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science*, 343(6167), pp.193–196.

Dixit, A. et al., 2016. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7), pp.1853–1866.e17.

Dobson, A.T. et al., 2004. The unique transcriptome through day 3 of human preimplantation development. *Human Molecular Genetics*, 13(14), pp.1461–1470.

Doench, J.G., 2018. Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews Genetics*, 19(2), pp.67–80.

Doench, J.G. et al., 2016. Optimised sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2), pp.184–191.

Doench, J.G. et al., 2014. Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nature Biotechnology*, 32, pp.1262

Doolittle, W.F. & Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature.*, 284(5757), pp.601–603.

Du, Z. et al., 2017. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature*, 547(7662), pp.232–235.

Eckersley-Maslin, M., Alda-Catalinas, C., Blotenburg, M., Kreibich, E., Krueger, C. & Reik, W., 2019. Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes & Development*, 33(3-4), pp.194-208

Eckersley-Maslin, M.A. et al., 2016. MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. *Cell Reports*, 17(1), pp.179–192.

Eckersley-Maslin, M.A., Alda-Catalinas, C. & Reik, W., 2018. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nature Reviews Molecular Cell Biology*, 19(7), pp.436–450.

Eden, E. et al., 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10, p.48.

Eid, A. et al., 2016. SUV4-20 activity in the preimplantation mouse embryo controls timely replication. *Genes & Development*, 30(22), pp.2513–2526.

Eng, C.-H.L. et al., 2019. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*, 568(7751), pp.235–239.

Eppig, J.J. & Schroeder, A.C., 1989. Capacity of mouse oocytes from preantral follicles to undergo embryogenesis and development to live young after growth, maturation, and fertilization in vitro. *Biology of Reproduction*, 41(2), pp.268–276.

Erhardt, S. et al., 2003. Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development. *Development*, 130(18), pp.4235–4248.

Erkek, S. et al., 2013. Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. *Nature Structural & Molecular Biology*, 20(7), pp.868–875.

Esposito, G. et al., 2007. Peptidylarginine deiminase (PAD) 6 is essential for oocyte cytoskeletal sheet formation and female fertility. *Molecular and Cellular Endocrinology*, 273(1-2), pp.25–31.

Evans, J.P., 1999. Sperm disintegrins, egg integrins, and other cell adhesion molecules of mammalian gamete plasma membrane interactions. *Front Biosci*, 4, pp.D114–31.

Evans, J.P. & Florman, H.M., 2002. The state of the union: the cell biology of fertilization. *Nature cell biology*, 4 Suppl, pp.s57–63.

Evans, M.J. & Kaufman, M.H., 1981. Establishment in culture of pluripotential cells from mouse embryos. *Nature.*, 292(5819), pp.154–156.

Fadloun, A. et al., 2013. Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nature Structural Molecular Biology*, 20(3), pp.332–338.

Falco, G. et al., 2007. Zscan4: A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Developmental Biology*, 307(2), pp.539–550.

Fan, H.C., Fu, G.K. & Fodor, S.P.A., 2015. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222), p.1258367.

Fan, X. et al., 2015. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome biology*, 16, p.148.

Faridani, O.R. et al., 2016. Single-cell sequencing of the small-RNA transcriptome. *Nature Biotechnology*, 34(12), pp.1264–1266.

Fehse, B. et al., 2004. Pois(s)on – It's a Question of Dose…. *Gene Therapy*, 11(11), pp.879–881.

Feldman, D. et al., 2018. Lentiviral co-packaging mitigates the effects of intermolecular recombination and multiple integrations in pooled genetic screens. *bioRxiv*, p.262121.

Ficz, G. et al., 2013. FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Stem Cell*, 13(3), pp.351–359.

Florman, H.M. & Wassarman, P.M., 1985. O-linked oligosaccharides of mouse egg ZP3 account for its sperm receptor activity. *Cell*, 41(1), pp.313–324.

Flyamer, I.M. et al., 2017. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648), pp.110–114.

Forlani, S. et al., 1998. Relief of a repressed gene expression state in the mouse 1-cell embryo requires DNA replication. *Development*, 125(16), p.3153.

Forrest, A.R.R. et al., 2014. A promoter-level mammalian expression atlas. *Nature*, 507(7493), pp.462–470.

Foygel, K. et al., 2008. A novel and critical role for Oct4 as a regulator of the maternal-embryonic transition. *PLoS ONE*, 3(12), p.e4109.

Franke, V. et al., 2017. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Research*, p.gr.216150.116.

Friedli, M. et al., 2014. Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome Research*, 24(8), pp.1251–1259.

Fu, B.X.H. et al., 2014. Landscape of target:guide homology effects on Cas9-mediated cleavage. *Nucleic Acids Research*, 42(22), pp.13778–13787.

Fu, X. et al., 2019. Myc and Dnmt1 impede the pluripotent to totipotent state transition in embryonic stem cells. *Nature cell biology*.

Gao, L. et al., 2018. Chromatin Accessibility Landscape in Human Early Embryos and Its Association with Evolution. *Cell*, 173(1), pp.248–259.e15.

Gao, Y. et al., 2017. Protein Expression Landscape of Mouse Embryos during Pre-implantation Development. *Cell Reports*, 21(13), pp.3957–3969.

Garcia, E. et al., 1999. RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1. *The EMBO journal*, 18(12), pp.3404–3418.

García-Pérez, J.L., Widmann, T.J. & Adams, I.R., 2016. The impact of transposable elements on mammalian development. *Development*, 143(22), p.4101.

Garneau, J.E. et al., 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, 468(7320), pp.67–71.

Gasiunas, G. et al., 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences*, 109(39), pp.E2579–86.

Gasperini, M. et al., 2019. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1), pp.377–390.e19.

Gasperini, M., Starita, L. & Shendure, J., 2016. The power of multiplexed functional analysis of genetic variants. *Nature Protocols*, 11(10), pp.1782–1787.

Gassler, J. et al., 2017. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *The EMBO journal*, 36(24), pp.3600–3618.

Gehring, J. et al., 2018. Highly Multiplexed Single-Cell RNA-seq for Defining Cell Population and Transcriptional Spaces. *bioRxiv*, p.315333.

Genga, R.M.J. et al., 2019. Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development. *Cell Reports*, 27(3), pp.708–718.e10.

Georgadaki, K. et al., 2016. The molecular basis of fertilization (Review). *International journal of molecular medicine*, 38(4), pp.979–986.

Georges, A. et al., 2014. FOXL2: a central transcription factor of the ovary. *Journal of molecular endocrinology*, 52(1), pp.R17–33.

Gierahn, T.M. et al., 2017. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, 14(4), pp.395–398.

Gilbert, L.A. et al., 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, 154(2), pp.442–451.

Gilbert, L.A. et al., 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell.*, 159(3), pp.647–661.

Ginsburg, M., Snow, M.H. & McLaren, A., 1990. Primordial germ cells in the mouse embryo during gastrulation. *Development*, 110(2), pp.521–528.

Graf, R. et al., 2019. sgRNA Sequence Motifs Blocking Efficient CRISPR/Cas9-Mediated Gene Editing. *Cell Reports*, 26(5), pp.1098–1103.e3.

Greally, J.M., 2018. A user's guide to the ambiguous word 'epigenetics'. *Nature Reviews Molecular Cell Biology*, 19(4), pp.207–208.

Greenberg, M.V.C. & Bourc'his, D., 2019. The diverse roles of DNA methylation in mammalian development and disease.

Grewal, S.I.S. & Elgin, S.C.R., 2002. Heterochromatin: new possibilities for the inheritance of structure. *Current Opinion in Genetics & Development*, 12(2), pp.178–187.

Gu, H., Zou, Y.R. & Rajewsky, K., 1993. Independent control of immunoglobulin switch recombination at individual switch regions evidenced through Cre-loxP-mediated gene targeting. *Cell*, 73(6), pp.1155–1164.

Gu, T.-P. et al., 2011. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature*, 477(7366), pp.606–610.

Guo, C. et al., 2018. CellTag Indexing: a genetic barcode-based multiplexing tool for single-cell technologies. *bioRxiv*, p.335547.

Guo, C. et al., 2019. CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome biology*, 20(1), p.90.

Guo, F. et al., 2014. Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell*, 15(4), pp.447–459.

Guo, F. et al., 2017. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Research*, 27(8), pp.967–988.

Guo, H. et al., 2014. The DNA methylation landscape of human early embryos. *Nature*, 511(7511), pp.606–610.

Guo, J. et al., 2017. An inducible CRISPR-ON system for controllable gene activation in human pluripotent stem cells. *Protein & Cell*, 8(5), pp.379–393.

Gurtu, V.E. et al., 2002. Maternal effect for DNA mismatch repair in the mouse. *Genetics*, 160(1), pp.271–277.

Habibi, E. et al., 2013. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell*, 13(3), pp.360–369.

Hadjantonakis, A.-K. & Arias, A.M., 2016. Single-Cell Approaches: Pandora's Box of Developmental Mechanisms. *Developmental Cell*, 38(6), pp.574–578.

Haeussler, M. et al., 2016. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome biology*, 17(1), p.148.

Hajkova, P. et al., 2010. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science*, 329(5987), pp.78–82.

Hamamoto, G. et al., 2014. Regulation of Transketolase Like 1 Gene Expression in the Murine One-Cell Stage Embryos A. J. Cooney, ed. *PLoS ONE*, 9(1), p.e82087.

Hamatani, T. et al., 2004. Dynamics of Global Gene Expression Changes during Mouse Preimplantation Development. *Developmental Cell*, 6(1), pp.117–131.

Hamazaki, N. et al., 2015. Gene activation-associated long noncoding RNAs function in mouse preimplantation development. *Development*, 142(5), pp.910–920.

Hanna, C.W. et al., 2018. MLL2 conveys transcription-independent H3K4 trimethylation in oocytes. *Nature Structural & Molecular Biology*, 25(1), pp.73–82.

Hara, K.T. et al., 2005. Cyclin A2-CDK2 regulates embryonic gene activation in 1-cell mouse embryos. *Developmental Biology*, 286(1), pp.102–113.

Harr, J.C., Gonzalez-Sandoval, A. & Gasser, S.M., 2016. Histones and histone modifications in perinuclear chromatin anchoring: from yeast to man. *EMBO reports*, 17(2), pp.139–155.

Hashimshony, T. et al., 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology*, 17, p.77.

Hashimshony, T. et al., 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports*, 2(3), pp.666–673.

Hatanaka, Y. et al., 2013. GSE is a maternal factor involved in active DNA demethylation in zygotes. *PLoS ONE*, 8(4), p.e60205.

Hatanaka, Y. et al., 2017. Histone H3 Methylated at Arginine 17 Is Essential for Reprogramming the Paternal Genome in Zygotes. *Cell Reports*, 20(12), pp.2756–2765.

Hayashi, K. et al., 2008. Dynamic Equilibrium and Heterogeneity of Mouse Pluripotent Stem Cells with Distinct Functional and Epigenetic States. *Stem Cell*, 3(4), pp.391–401.

Hayashi, M. et al., 2015. Chd5 Regulates MuERV-L/MERVL Expression in Mouse Embryonic Stem Cells Via H3K27me3 Modification and Histone H3.1/H3.2. *Journal of Cellular Biochemistry*, 117(3), pp.780–792.

Hayashi, T. et al., 2018. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nature Communications*, 9(1), p.619.

Heler, R. et al., 2015. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature*, 519(7542), pp.199–202.

Hendrickson, P.G. et al., 2017. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nature Genetics*, 49(6), pp.925-934

Henery, C.C. et al., 1995. Repression of gene expression at the beginning of mouse development. *Developmental Biology*, 169(2), pp.448–460.

Herbst, F. et al., 2012. Extensive Methylation of Promoter Sequences Silences Lentiviral Transgene Expression During Stem Cell Differentiation In Vivo. *Molecular Therapy*, 20(5), pp.1014–1021.

Hermans, P.W. et al., 1991. Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. *Infection and immunity*, 59(8), pp.2695–2705.

Hernandez, C. et al., 2018. Dppa2/4 Facilitate Epigenetic Remodeling during Reprogramming to Pluripotency. *Cell Stem Cell*, 23(3), pp.396–411.e8.

Hess, G.T. et al., 2016. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nature Methods*, 13(12), pp.1036–1042.

Hill, A.J. et al., 2018. On the design of CRISPR-based single-cell molecular screens. *Nature Methods*, 16, p.299.

Hilton, I.B. et al., 2015. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, 33(5), pp.510–517.

Hirata, T. et al., 2012. Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells. *Science* Reports, 2, pp,208.

Hisada, K. et al., 2012. RYBP Represses Endogenous Retroviruses and Preimplantation- and Germ Line-Specific Genes in Mouse Embryonic Stem Cells. *Molecular and Cellular Biology*, 32(6), pp.1139–1149.

Holliday, R. & Pugh, J.E., 1975. DNA modification mechanisms and gene activity during development. *Science*, 187(4173), pp.226–232.

Horlbeck, M.A., Gilbert, L.A., Villalta, J.E., Adamson, B., Pak, R.A., Chen, Y., Fields, A.P., Park, C.Y., Corn, J.E., Kampmann, M. & Weissman, J.S., 2016. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation K. Adelman, ed. *eLife*, 5, p.e19760.

Horlbeck, M.A., Witkowsky, L.B., et al., 2016. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife*, 5.

Hota, S.K. & Bruneau, B.G., 2016. ATP-dependent chromatin remodeling during mammalian development. *Development*, 143(16), p.2882.

Howell, C.Y. et al., 2001. Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell.*, 104(6), pp.829–838.

Hsu, P.D. et al., 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, 31, pp.827

Hu, J. et al., 2010. Mouse ZAR1-like (XM_359149) colocalizes with mRNA processing components and its dominant-negative mutant caused two-cell-stage embryonic arrest. *Developmental dynamics: an official publication of the American Association of Anatomists*, 239(2), pp.407–424.

Huang, L. et al., 2019. Rad9a is involved in chromatin decondensation and post-zygotic embryo development in mice. *Cell Death Differentation*, 26(5), pp.968-980

Huang, X.-T. et al., 2018. Technical Advances in Single-Cell RNA Sequencing and Applications in Normal and Malignant Hematopoiesis. *Frontiers in Oncology*, 8, p.582.

Huang, Y. et al., 2017. STELLA modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *eLife*, 6, p.e22345.

Iaco, A.D. et al., 2019. DUX is a non-essential synchronizer of zygotic genome activation. *bioRxiv*, p.569434.

Imbeault, M., Helleboid, P.-Y. & Trono, D., 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646), pp.550–554.

Inoue, A. & Zhang, Yi, 2014. Nucleosome assembly is required for nuclear pore complex assembly in mouse zygotes. *Nature Structural & Molecular Biology*, 21(7), pp.609–616.

Inoue, A. & Zhang, Yi, 2011. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science*, 334(6053), p.194.

Inoue, A. et al., 2015. Haploinsufficiency, but not defective paternal 5mC oxidation, accounts for the developmental defects of maternal Tet3 knockouts. *Cell Reports*, 10(4), pp.463–470.

Inoue, A. et al., 2017. Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature*, 547(7664), pp.419–424.

Inoue, A., Matoba, S. & Zhang, Y., 2012. Transcriptional activation of transposable elements in mouse zygotes is independent of Tet3-mediated 5-methylcytosine oxidation. *Cell Research*, 22(12), pp.1640–1649.

Iqbal, K. et al., 2011. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National Academy of Sciences*, 108(9), pp.3642–3647.

Isaac, R.S. et al., 2016. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *eLife*, 5.

Ishiguro, K.-I., Monti, M., et al., 2016. Zscan4 is expressed specifically during late meiotic prophase in both spermatogenesis and oogenesis. *In Vitro Cellular & Developmental Biology - Animal*, pp.1–12.

Ishiguro, K.-I., Nakatake, Y., et al., 2016. Expression analysis of the endogenous Zscan4 locus and its coding proteins in mouse ES cells and preimplantation embryos. *In Vitro Cellular & Developmental Biology - Animal*, pp.1–12.

Ishino, Y. et al., 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of bacteriology*, 169(12), pp.5429–5433.

Ishiuchi, T. & Torres-Padilla, M.-E., 2013. Towards an understanding of the regulatory mechanisms of totipotency. *Current Opinion in Genetics & Development*, 23(5), pp.512–518.

Ishiuchi, T. et al., 2015. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nature Structural & Molecular Biology*, pp.1–13.

Islam, S. et al., 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7), pp.1160–1167.

Islam, S. et al., 2012. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nature Protocols*, 7(5), pp.813–828.

Islam, S. et al., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2), pp.163–166.

Iturbide, A. & Torres-Padilla, M.-E., 2017. Starting embryonic transcription for the first time. *Nature Genetics*, 49(6), pp.820–821.

Ivanova, I. et al., 2017. The RNA m(6)A Reader YTHDF2 Is Essential for the Post-transcriptional Regulation of the Maternal Transcriptome and Oocyte Competence. *Molecular Cell*, 67(6), pp.1059–1067.e4.

Jachowicz, J.W. et al., 2013. Heterochromatin establishment at pericentromeres depends on nuclear position. *Genes & Development*, 27(22), pp.2427–2432.

Jachowicz, J.W. et al., 2017. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nature Genetics*, 49(10), pp.1502–1510.

Jaitin, D.A. et al., 2016. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, 167(7), pp.1883–1896.e15.

Jaitin, D.A. et al., 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172), pp.776–779.

Jansen, R. et al., 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular microbiology*, 43(6), pp.1565–1575.

Jansz, N. & Torres-Padilla, M.-E., 2019. Genome activation and architecture in the early mammalian embryo. *Current Opinion in Genetics & Development*, 55, pp.52–58.

Jensen, K.T. et al., 2017. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS letters*, 591(13), pp.1892–1901.

Jiang, J. et al., 2013. Zscan4 promotes genomic stability during reprogramming and dramatically improves the quality of iPS cells as demonstrated by tetraploid complementation. *Cell Research*, 23(1), pp.92–106.

Jinek, M. et al., 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096), pp.816–821.

Johnson, M.H. & Ziomek, C.A., 1981. The foundation of two distinct cell lineages within the mouse morula. *Cell*, 24(1), pp.71–80.

Joseph, A., Mitchell, A.R. & Miller, O.J., 1989. The organization of the mouse satellite DNA at centromeres. *Experimental cell research*, 183(2), pp.494–500.

Jouhilahti, E.-M. et al., 2016. The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation. *Development*, 143(19), pp.3459–3469.

Joung, J. et al., 2017. Genome-scale CRISPR-Cas9 Knockout and Transcriptional Activation Screening, *Nature Protocols*, 12(4), pp.828-863.

Jukam, D., Shariati, S.A.M. & Skotheim, J.M., 2017. Zygotic Genome Activation in Vertebrates. *Developmental Cell*, 42(4), pp.316–332.

Jung, Y.H. et al., 2017. Chromatin States in Mouse Sperm Correlate with Embryonic and Adult Regulatory Landscapes. *Cell Reports*, 18(6), pp.1366–1382.

Kaneda, M. et al., 2009. Essential role for Argonaute2 protein in mouse oogenesis. *Epigenetics & chromatin*, 2(1), p.9.

Kaneda, M. et al., 2004. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting., 429(6994), pp.900–903.

Kang, H.M. et al., 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36, pp.89-94.

Karlic, R. et al., 2017. Long non-coding RNA exchange during the oocyte-to-embryo transition in mice. *DNA Research*, 24(2), pp.129-141

Kaya-Okur, H.S. et al., 2019. "CUT&Tag for efficient epigenomic profiling of small samples and single cells." *Nature communications,* 10(1) p.1930.

Ke, Y. et al., 2017. 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell*, 170(2), pp.367–381.e20.

Kearns, N.A. et al., 2015. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nature Methods*, 12(5), pp.401–403.

Kelsey, G., Stegle, O. & Reik, W., 2017. Single-cell epigenomics: Recording the past and predicting the future. *Science*, 358(6359), pp.69–75.

Kigami, D., 2002. MuERV-L Is One of the Earliest Transcribed Genes in Mouse One-Cell Embryos. *Biology of Reproduction*, 68(2), pp.651–654.

Kim, D. et al., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4), p.R36.

Kim, D., Langmead, B. & Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4), pp.357–360.

Kim, J. et al., 2016. Maternal Setdb1 Is Required for Meiotic Progression and Preimplantation Development in Mouse. M.-H. Verlhac, ed. *PLOS Genetics*, 12(4), p.e1005970.

Kim, K.-H. & Lee, K.-A., 2014. Maternal effect genes: Findings and effects on mouse embryo development. *Clinical and experimental reproductive medicine*, 41(2), pp.47–61.

Kim, K.-H., Seo, Y.-M., Kim, E.-Y., Lee, S.-Y., Kwon, J., Ko, J.-J. & Lee, K.-A., 2016. The miR-125 family is an important regulator of the expression and maintenance of maternal

effect genes during preimplantational embryo development. *Open biology*, 6(11), poo:160181

King, H.W. et al., 2018. Polycomb repressive complex 1 shapes the nucleosome landscape but not accessibility at target genes. *Genome Research*, 28(10), pp.1494–1507.

Klein, A.M. et al., 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5), pp.1187–1201.

Klose, R.J. & Bird, A.P., 2006. Genomic DNA methylation: the mark and its mediators. *Trends in biochemical sciences*, 31(2), pp.89–97.

Knight, S.C. et al., 2015. Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science*, 350(6262), p.823.

Kobayashi, H. et al., 2012. Contribution of Intragenic DNA Methylation in Mouse Gametic DNA Methylomes to Establish Oocyte-Specific Heritable Marks W. Reik, ed. *PLoS Genet.*, 8(1), p.e1002440.

Koike-Yusa, H. et al., 2014. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, 32(3), pp.267–273.

Konermann, S. et al., 2015. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, 517(7536), pp.583–588.

Konermann, S. et al., 2013. Optical control of mammalian endogenous transcription and epigenetic states. *Nature*, 500(7463), pp.472–476.

Kouzarides, T., 2007. Chromatin modifications and their function. *Cell*, 128(4), pp.693–705.

Krauchunas, A.R. & Wolfner, M.F., 2013. Molecular changes during egg activation. *Current topics in developmental biology*, 102, pp.267–292.

Krueger, F. & Andrews, S.R., 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11), pp.1571–1572.

Kruse, K. et al., 2019. Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. *bioRxiv*, p.523712.

Kumar, T.R. et al., 1997. Follicle stimulating hormone is required for ovarian follicle maturation but not male fertility. *Nature Genetics*, 15(2), pp.201–204.

Kunarso, G. et al., 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, 42(7), pp.631–634.

Kuscu, C. et al., 2014. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nature Biotechnology*, 32(7), pp.677–683.

Kwon, D.Y. et al., 2017. Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. *Nature Communications*, 8, p.15315.

Kwon, Y.-W. et al., 2015. Role of Zscan4 in secondary murine iPSC derivation mediated by protein extracts of ESC or iPSC. *Biomaterials*, 59, pp.102–115.

La Fuente, De, R. et al., 2004. Major chromatin remodeling in the germinal vesicle (GV) of mammalian oocytes is dispensable for global transcriptional silencing but required for centromeric heterochromatin function. *Developmental Biology*, 275(2), pp.447–458.

La Russa, M.F. & Qi, L.S., 2015. The New State of the Art: Cas9 for Gene Activation and Repression. *Mol Cell Biol.*, 35(22), p.3800.

Labuhn, M. et al., 2017. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR–Cas9 applications. *Nucleic Acids Research*, 46(3), pp.1375–1385.

Ladstatter, S. & Tachibana, K., 2019. Genomic insights into chromatin reprogramming to totipotency in embryos. *J Cell Biol*, 218(1), pp.70–82.

Ladstatter, S. & Tachibana-Konwalski, K., 2016. A Surveillance Mechanism Ensures Repair of DNA Lesions during Zygotic Reprogramming. *Cell*, 167(7), pp.1774-1787.e13

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), pp.357–359.

Larson, M.H. et al., 2013. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols*, 8(11), pp.2180–2196.

Lawrence, M., Daujat, S. & Schneider, R., 2016. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics*, 32(1), pp.42–56.

Leader, B. et al., 2002. Formin-2, polyploidy, hypofertility and positioning of the meiotic spindle in mouse oocytes. *Nature cell biology*, 4(12), pp.921–928.

Lee, H.J., Hore, T.A. & Reik, W., 2014. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell*, 14(6), pp.710–719.

Lee, M.T., Bonneau, A.R. & Giraldez, A.J., 2014. Zygotic Genome Activation During the Maternal-to-Zygotic Transition. *Annual Review of Cell and Developmental Biology*, 30(1), pp.581–613.

Leidenroth, A. et al., 2012. Evolution of DUX gene macrosatellites in placental mammals. *Chromosoma*, 121(5), pp.489–497.

Lepikhov, K. & Walter, J., 2004. Differential dynamics of histone H3 methylation at positions K4 and K9 in the mouse zygote. *BMC Developmental Biology*, 4(1), p.12.

Leung, C.Y. & Zernicka-Goetz, M., 2015. Mapping the journey from totipotency to lineage specification in the mouse embryo. *Current Opinion in Genetics & Development*, 34, pp.71–76.

Lewandoski, M., Wassarman, K.M. & Martin, G.R., 1997. Zp3-cre, a transgenic mouse line for the activation or inactivation of loxP-flanked target genes specifically in the female germ line. *Current biology: CB*, 7(2), pp.148–151.

Lewis, C.J.T., Pan, T. & Kalsotra, A., 2017. RNA modifications and structures cooperate to guide RNA-protein interactions. *Nature Reviews Molecular Cell Biology*, 18(3), pp.202–210.

Li, Dongwei et al., 2017. Chromatin Accessibility Dynamics during iPSC Reprogramming. *Cell Stem Cell*, 21(6), pp.819–833.e6.

Li, E, Bestor, T.H. & Jaenisch, R., 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6), pp.915–926.

Li, Heng & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Li, Heng et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Li, Lei, Baibakov, B. & Dean, J., 2008. A Subcortical Maternal Complex Essential for Preimplantation Mouse Embryogenesis. *Developmental Cell*, 15(3), pp.416–425.

Li, Lei, Lu, X. & Dean, J., 2013. The maternal to zygotic transition in mammals. *Molecular Aspects of Medicine*, 34(5), pp.919–938.

Li, Lei, Zheng, P. & Dean, J., 2010. Maternal control of early mouse development. *Development*, 137(6), pp.859–870.

Li, Lin et al., 2018. Single-cell multi-omics sequencing of human early embryos. *Nature cell biology*, 20(7), pp.847–858.

Li, Rong & Albertini, D.F., 2013. The road to maturation: somatic cell interaction and self-organization of the mammalian oocyte. *Nature Reviews Molecular Cell Biology*, 14(3), pp.141–152.

Li, Xiajun et al., 2008. A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. *Developmental Cell*, 15(4), pp.547–557.

Li, Yingfeng et al., 2018. Stella safeguards the oocyte methylome by preventing de novo methylation mediated by DNMT1. *Nature*, 564(7734), pp.136-140

Liang, G. et al., 2004. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19), p.7357.

Lin, C.-J. et al., 2014. Hira-Mediated H3.3 Incorporation Is Required for DNA Replication and Ribosomal RNA Transcription in the Mouse Zygote. *Developmental Cell*, 30(3), pp.268–279.

Lin, K. et al., 2018. Essential requirement of mammalian Pumilio family in embryonic development. *Molecular Biology of the Cell*, 29(24), pp.2922–2932.

Lino, C.A. et al., 2018. Delivering CRISPR: a review of the challenges and approaches. *Drug delivery*, 25(1), pp.1234–1257.

Liu, Peng et al., 2018. CRISPR-Based Chromatin Remodeling of the Endogenous Oct4 or Sox2 Locus Enables Reprogramming to Pluripotency. *Cell Stem Cell*, 22(2), pp.252-261.

Liu, Serena & Trapnell, C., 2016. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research*, 5.

Liu, Xiaoyu et al., 2016. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature*, 537(7621), pp.558–562.

Liu, Yang et al., 2019. Maternal DCAF13 Regulates Chromatin Tightness to Contribute to Embryonic Development. *Scientific Reports*, 9(1), p.6278.

Liu, Yanxia et al., 2018. CRISPR Activation Screens Systematically Identify Factors that Drive Neuronal Fate and Reprogramming. *Cell Stem Cell*, 23(5), pp.758–771.e8.

Liu, Yusheng et al., 2016. BTG4 is a key regulator for maternal mRNA clearance during mouse early embryogenesis. *Journal of molecular cell biology*, 8(4), pp.366–368.

Lo, A. & Qi, L., 2017. Genetic and epigenetic control of gene expression by CRISPR-Cas systems. *F1000Research*, 6.

Local, A. et al., 2018. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nature Genetics*, 50(1), pp.73–82.

Lu, F., Liu, Y., Inoue, A., Suzuki, T., Zhao, K. & Zhang, Y., 2016. Establishing Chromatin Regulatory Landscape during Mouse Preimplantation Development. *Cell.*, 165(6), pp.1375–1388.

Lu, X. et al., 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology*, 21(4), pp.423–425.

Lubeck, E. et al., 2014. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*, 11(4), pp.360–361.

Luo, Y. et al., 2018. Integrative single-cell omics analyses reveal epigenetic heterogeneity in mouse embryonic stem cells. *PLoS computational biology*, 14(3), p.e1006034.

Lv, J. et al., 2015. Identification of 4438 novel lincRNAs involved in mouse pre-implantation embryonic development. *Molecular genetics and genomics: MGG*, 290(2), pp.685–697.

Lykke-Andersen, K. et al., 2008. Maternal Argonaute 2 is essential for early mouse development at the maternal-zygotic transition. *Molecular Biology of the Cell*, 19(10), pp.4383–4392.

Ma, H. et al., 2014. Pol III Promoters to Express Small RNAs: Delineation of Transcription Initiation. *Molecular Therapy - Nucleic Acids*, 3, p.e161.

Ma, Y. et al., 2016. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nature Methods*, 13(12), pp.1029–1035.

Macfarlan, T.S. et al., 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature.*, 487 (7405), pp.57-63.

Macfarlan, T.S. et al., 2011. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes & Development*, 25(6), pp.594–607.

Macosko, E.Z. et al., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), pp.1202–1214.

Madissoon, E. et al., 2019. Pleomorphic Adenoma Gene 1 Is Needed For Timely Zygotic Genome Activation and Early Embryo Development. *Scientific Reports*, 9(1), p.8411.

Maeder, M.L. et al., 2013. CRISPR RNA–guided activation of endogenous human genes. *Nature Methods*, 10(10), pp.977–979.

Maenohara, S. et al., 2017. Role of UHRF1 in de novo DNA methylation in oocytes and maintenance methylation in preimplantation embryos. *PLOS Genetics*, 13(10), p.e1007042.

Mager, D.L. & Stoye, J.P., 2015. Mammalian Endogenous Retroviruses. *Microbiology spectrum*, 3(1), pp.MDNA3–0009–2014.

Mahadevan, S. et al., 2017. Maternally expressed NLRP2 links the subcortical maternal complex (SCMC) to fertility, embryogenesis and epigenetic reprogramming. *Scientific Reports*, 7, p.44667.

Majumder, S. & DePamphilis, M.L., 1995. A unique role for enhancers is revealed during early mouse development. *BioEssays*, 17(10), pp.879–889.

Majumder, S., Miranda, M. & DePamphilis, M.L., 1993. Analysis of gene expression in mouse preimplantation embryos demonstrates that the primary role of enhancers is to relieve repression of promoters. *The EMBO journal*, 12(3), pp.1131–1140.

Mak, W. et al., 2018. A role of Pumilio 1 in mammalian oocyte maturation and maternal phase of embryogenesis. *Cell & bioscience*, 8, p.54.

Makarova, K.S. et al., 2015. An updated evolutionary classification of CRISPR-Cas systems. *Nature reviews. Microbiology*, 13(11), pp.722–736.

Maksakova, I.A. et al., 2013. Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. *Epigenetics & chromatin*, 6(1), pp.1–16.

Mali, P., Aach, J., et al., 2013. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnology*, 31(9), pp.833–838.

Mali, P., Yang, L., et al., 2013. RNA-guided human genome engineering via Cas9. *Science*, 339(6121), pp.823–826.

Martin, G.R., 1981. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12), pp.7634–7638.

Maurer, J. et al., 2008. Contrasting expression of keratins in mouse and human embryonic stem cells. *PLoS ONE*, 3(10), pp.e3451–e3451.

Mayer, W. et al., 2000. Demethylation of the zygotic paternal genome. *Nature*, 403(6769), pp.501–502.

McGinnis, C.S., Murrow, L.M. & Gartner, Z.J., 2019. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, 8(4), pp.329–337.e4.

McInnes, L., Healy, J. & Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Medvedev, S. et al., 2008. CDC2A (CDK1)-mediated phosphorylation of MSY2 triggers maternal mRNA degradation during mouse oocyte maturation. *Developmental Biology*, 321(1), pp.205–215.

Medvedev, S., Pan, H. & Schultz, R.M., 2011. Absence of MSY2 in Mouse Oocytes Perturbs Oocyte Growth and Maturation, RNA Stability, and the Transcriptome1. *Biology of Reproduction*, 85(3), pp.575–583.

Meyenn, von, F. et al., 2016. Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells. *Molecular Cell*, 62(6), pp.848–861.

Mi, H. et al., 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47(D1), pp.D419–D426.

Miao, Y.-L. et al., 2018. Mediator complex component MED13 regulates zygotic genome activation and is required for postimplantation development in the mouse. *Biology of Reproduction*, 98(4), pp.449-464.

Milagre, I. et al., 2017. Gender Differences in Global but Not Targeted Demethylation in iPSC Reprogramming. *CellReports*, 18(5), pp.1079–1089.

Mishima, Y. & Tomari, Y., 2016. Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish. *Molecular Cell*, 61(6), pp.874–885.

Moffitt, J.R. & Zhuang, X., 2016. RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). *Methods in enzymology*, 572, pp.1–49.

Moffitt, J.R. et al., 2016. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39), pp.11046–11051.

Mohr, S.E. & Perrimon, N., 2012. RNAi screening: new approaches, understandings, and organisms. *Wiley interdisciplinary reviews. RNA*, 3(2), pp.145–158.

Mojica, F.J., Juez, G. & Rodriguez-Valera, F., 1993. Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Molecular microbiology*, 9(3), pp.613–621.

Mojica, F.J.M. et al., 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of molecular evolution*, 60(2), pp.174–182.

Mojica, F.J.M. et al., 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, England)*, 155(Pt 3), pp.733–740.

Morgan, M. et al., 2017. mRNA 3′ uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. *Nature*, 548(7667), pp.347–351.

Morgani, S.M. et al., 2013. Totipotent Embryonic Stem Cells Arise in Ground-State Culture Conditions. *Cell Reports*, 3(6), pp.1945–1957.

Mtango, N.R. et al., 2012. Essential role of maternal UCHL1 and UCHL3 in fertilization and preimplantation embryo development. *Journal of Cellular Physiology*, 227(4), pp.1592–1603.

Mulholland, C.B. et al., 2018. TET1 drives global DNA demethylation via DPPA3-mediated inhibition of maintenance methylation. *bioRxiv*, p.321604.

Murchison, E.P. et al., 2007. Critical roles for Dicer in the female germline. *Genes & Development*, 21(6), pp.682–693.

Nakamura, T. et al., 2012. PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. *Nature*, 486(7403), pp.415–419.

Nashun, B. et al., 2015. Continuous Histone Replacement by Hira Is Essential for Normal Transcriptional Regulation and De Novo DNA Methylation during Mouse Oogenesis. *Molecular Cell*, 60(4), pp.611–625.

Neri, F. et al., 2017. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, 543(7643), pp.72–77.

Nonchev, S. & Tsanev, R., 1990. Protamine-histone replacement and DNA replication in the male mouse pronucleus. *Mol Reprod Dev*, 25(1), pp.72–76.

Norman, T.M. et al., 2019. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*.

Nothias, J.Y., Miranda, M. & DePamphilis, M.L., 1996. Uncoupling of transcription and translation during zygotic gene activation in the mouse. *The EMBO journal*, 15(20), pp.5715–5725.

O'Carroll, D. et al., 2001. The polycomb-group gene Exh2 is Required for Early Mouse Development. *Mol Cell Biol.*, 21(13), p.4330-6

Ohno, S., 1972. So much "junk" DNA in our genome. *Brookhaven symposia in biology*, 23, pp.366–370.

Okamura, K. & Nakai, K., 2008. Retrotransposition as a source of new promoters. *Molecular biology and evolution*, 25(6), pp.1231–1238.

Okano, M. et al., 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3), pp.247–257.

Olins, D.E. & Olins, A.L., 2003. Chromatin history: our view from the bridge. *Nature Reviews Molecular Cell Biology*, 4(10), pp.809–814.

Ong, C.-T. & Corces, V.G., 2014. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics*, 15(4), pp.234–246.

Oqani, R.K. et al., 2019. Iws1 and Spt6 Regulate Trimethylation of Histone H3 on Lysine 36 through Akt Signaling and are Essential for Mouse Embryonic Genome Activation. *Scientific* Reports, 9(1), pp.3831–3831.

Orgel, L.E. & Crick, F.H.C., 1980. Selfish DNA: the ultimate parasite. *Nature.*, 284(5757), pp.604–607.

Oswald, J. et al., 2000. Active demethylation of the paternal genome in the mouse zygote. *Current biology: CB*, 10(8), pp.475–478.

Ow, J.R. et al., 2014. Patz1 regulates embryonic stem cell identity. *Stem cells and development*, 23(10), pp.1062–1073.

Pan, H. & Schultz, R.M., 2011. Sox2 modulates reprogramming of gene expression in two-cell mouse embryos. *Biology of Reproduction*, 85(2), pp.409–416.

Pan, H. et al., 2005. Transcript profiling during mouse oocyte development and the effect of gonadotropin priming and development in vitro. *Developmental Biology*, 286(2), pp.493–506.

Parekh, U. et al., 2018. Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. *Cell Systems*, 7(5), pp.548–555.e8.

Park, S.-J. et al., 2015. DBTMEE: a database of transcriptome in mouse early embryos. *Nucleic Acids Research*, 43(Database issue), pp.D771–6.

Payer, B. et al., 2003. Stella Is a Maternal Effect Gene Required for Normal Early Development in Mice. *Current Biology*, 13(23), pp.2110–2117.

Paynton, B.V. & Bachvarova, R., 1994. Polyadenylation and deadenylation of maternal mRNAs during oocyte growth and maturation in the mouse. *Mol Reprod Dev*, 37(2), pp.172–180.

Pearson, K., 2010. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), pp.559–572.

Peaston, A.E. et al., 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Developmental Cell*, 7(4), pp.597–606.

Peat, J.R. et al., 2014. Genome-wide Bisulfite Sequencing in Zygotes Identifies Demethylation Targets and Maps the Contribution of TET3 Oxidation. *Cell Reports*, 9(6), pp.1990–2000.

Peng, G. et al., 2019. Molecular architecture of lineage allocation and tissue organization in early mouse embryo. *Nature*, 572(7770), pp.528-532.

Penzkofer, T. et al., 2017. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Research*, 45(D1), pp.D68–D73.

Percharde, M. et al., 2018. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell*, 174(2), pp.191-405.e19

Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., Guilak, F., Crawford, G.E.,

Reddy, T.E. & Gersbach, C.A., 2013. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature Methods*, 10(10), pp.973–976.

Petrussa, L., van de Velde, H. & De Rycke, M., 2016. Similar kinetics for 5-methylcytosine and 5-hydroxymethylcytosine during human preimplantation development in vitro. *Mol Reprod Dev*, 83(7), pp.594–605.

Pfeifer, A. et al., 2002. Transgenesis by lentiviral vectors: lack of gene silencing in mammalian embryonic stem cells and preimplantation embryos. *Proceedings of the National Academy of Sciences of the United States of America*, 99(4), pp.2140–2145.

Pfeiffer, M.J. et al., 2015. Differences in embryo quality are associated with differences in oocyte composition: A proteomic study in inbred mice. *Proteomics*, 15(4), pp.675–687.

Pfeiffer, M.J. et al., 2011. Proteomic Analysis of Mouse Oocytes Reveals 28 Candidate Factors of the "Reprogrammome." *Journal of Proteome Research*, 10(5), pp.2140–2153.

Pfender, S. et al., 2015. Live imaging RNAi screen reveals genes essential for meiosis in mammalian oocytes. *Nature.*, 524(7564), pp.239–242.

Philipps, D.L. et al., 2008. The dual bromodomain and WD repeat-containing mouse protein BRWD1 is required for normal spermiogenesis and the oocyte-embryo transition. *Developmental Biology*, 317(1), pp.72–82.

Picelli, S. et al., 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1), pp.171–181.

Picelli, S. et al., 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11), pp.1096–1098.

Pickar-Oliver, A. & Gersbach, C.A., 2019. The next generation of CRISPR-Cas technologies and applications. *Nature Reviews Molecular Cell Biology*, 20(8), pp.490–507.

Pijuan-Sala, B. et al., 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745), pp.490–495.

Piotrowska-Nitsche, K. et al., 2005. Four-cell stage mouse blastomeres have different developmental properties. *Development*, 132(3), p.479.

Polstein, L.R. et al., 2015. Genome-wide specificity of DNA binding, gene regulation, and chromatin remodeling by TALE- and CRISPR/Cas9-based transcriptional activators. *Genome Research*, 25(8), pp.1158–1169.

Pontis, J. et al., 2019. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell*, 24(5), pp.724-735.e5.

Portela, A. & Esteller, M., 2010. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10), pp.1057–1068.

Posfai, E. et al., 2012. Polycomb function during oogenesis is required for mouse embryonic development. *Genes & Development*, 26(9), pp.920–932.

Pradeepa, M.M. et al., 2016. Histone H3 globular domain acetylation identifies a new class of enhancers. *Nature Genetics*, 48(6), pp.681–686.

Probst, A.V. et al., 2010. A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. *Developmental Cell*, 19(4), pp.625–638.

Probst, A.V. et al., 2007. Structural differences in centromeric heterochromatin are spatially reconciled on fertilisation in the mouse zygote. *Chromosoma*, 116(4), pp.403–415.

Qi, L.S. et al., 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152(5), pp.1173–1183.

Qiu, J.-J., Ren, Z.-R. & Yan, J.-B., 2016. Identification and functional analysis of long non-coding RNAs in human and mouse early embryos based on single-cell transcriptome data. *Oncotarget*, 7(38), pp.61215-61228.

Qiu, P., 2018. Embracing the dropouts in single-cell RNA-seq data. *bioRxiv*, p.468025.

Quenneville, S. et al., 2011. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Molecular Cell*, 44(3), pp.361–372.

Racki, W.J. & Richter, J.D., 2006. CPEB controls oocyte growth and follicle development in the mouse. *Development*, 133(22), p.4527.

Rada-Iglesias, A. et al., 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333), pp.279–283.

Radzisheuskaya, A. et al., 2016. Optimizing sgRNA position markedly improves the efficiency of CRISPR/dCas9-mediated transcriptional repression. *Nucleic Acids Research*, 44(18), pp.e141–e141.

Rajagopal, N. et al., 2016. High-throughput mapping of regulatory DNA. *Nature Biotechnology*, 34(2), pp.167–174.

Ram, P.T. & Schultz, R.M., 1993. Reporter gene expression in G2 of the 1-cell mouse embryo. *Developmental Biology*, 156(2), pp.552–556.

Ramos, S.B.V. et al., 2004. The CCCH tandem zinc-finger protein Zfp36l2 is crucial for female fertility and early embryonic development. *Development*, 131(19), pp.4883–4893.

Ramsköld, D. et al., 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8), pp.777–782.

Razin, A. & Riggs, A.D., 1980. DNA methylation and gene function. *Science*, 210(4470), pp.604–610.

Regev, A. et al., 2017. The Human Cell Atlas. *eLife*, 6.

Reik, W. & Walter, J., 2001. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1), pp.21–32.

Reik, W., Santos, F. & Dean, W., 2003. Mammalian epigenomics: reprogramming the genome for development and therapy. *Theriogenology*, 59(1), pp.21–32.

Replogle, J.M., Xu, A., Norman, T.M., Meer, E.J., Terry, J.M., Riordan, D., Srinivas, N., Mikkelsen, T.S., Weissman, J.S. & Adamson, B., 2018. Direct capture of CRISPR guides enables scalable, multiplexed, and multi-omic Perturb-seq. *bioRxiv*, p.503367.

Richardson, S.R. et al., 2015. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology spectrum*, 3(2), pp.MDNA3–2014.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), pp.139–140.

Rodriguez-Terrones, D. & Torres-Padilla, M.-E., 2018. Nimble and Ready to Mingle: Transposon Outbursts of Early Development. *Trends in Genetics*, 34(10), pp.806-820.

Rodriguez-Terrones, D. et al., 2018. A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nature Genetics*, 50(1), pp.106–119.

Rodriques, S.G. et al., 2019. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434), pp.1463–1467.

Roest, H.P. et al., 2004. The ubiquitin-conjugating DNA repair enzyme HR6A is a maternal factor essential for early embryonic development in mice. *Molecular and Cellular Biology*, 24(12), pp.5485–5495.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., Pun, S.H., Sellers, D.L., Tasic, B. & Seelig, G., 2018. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385), pp.176–182.

Rossant, J., 1976. Postimplantation development of blastomeres isolated from 4- and 8-cell mouse eggs. *Journal of embryology and experimental morphology*, 36(2), pp.283–290.

Rossant, J. & Tam, P.P.L., 2009. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development*, 136(5), pp.701–713.

Rougier, N. et al., 1998. Chromosome methylation patterns during mammalian preimplantation development. *Genes & Development*, 12(14), pp.2108–2113.

Rowe, H.M. & Trono, D., 2011. Dynamic control of endogenous retroviruses during development. *Virology*, 411(2), pp.273–287.

Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P.V., Layard-Liesching, H., Verp, S., Marquis, J., Spitz, F., Constam, D.B. & Trono, D., 2010. KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, 463(7278), pp.237–240.

Rubin, A.J. et al., 2019. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell*, 176(1-2), pp.361–376.e17.

Rulands, S. et al., 2018. Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency. *Cell Systems*, 7(1), pp.63–76.e12.

Russell, D.L. & Robker, R.L., 2007. Molecular mechanisms of ovulation: co-ordination through the cumulus complex. *Human Reproduction Update*, 13(3), pp.289–312.

Salmen, F. et al., 2018. Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. *Nature Protocols*, 13(11), pp.2501–2534.

Sander, J.D. & Joung, J.K., 2014. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, 32(4), pp.347–355.

Sanson, K.R. et al., 2018. Optimised libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, 9(1), p.5416.

Santenard, A. et al., 2010. Heterochromatin formation in the mouse embryo requires critical residues of the histone variant H3.3. *Nature Publishing Group*, 12(9), pp.853–862.

Santos, F. et al., 2013. Active demethylation in mouse zygotes involves cytosine deamination and base excision repair. *Epigenetics & chromatin*, 6(1), p.39.

Santos, F. et al., 2005. Dynamic chromatin modifications characterise the first cell cycle in mouse embryos. *Developmental Biology*, 280(1), pp.225–236.

Sasagawa, Y. et al., 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology*, 14(4), p.R31.

Satija, R. et al., 2015. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, *33*(5), 495–502.

Satpathy, A.T. et al., 2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, 37(8), pp.925–936.

Schaum, N. et al., 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature.*, 562(7727), pp.367–372.

Schlesinger, S. & Goff, S.P., 2015. Retroviral Transcriptional Regulation and Embryonic Stem Cells: War and Peace. *Molecular and Cellular Biology*, 35(5), pp.770–777.

Schoorlemmer, J. et al., 2014. Regulation of Mouse Retroelement MuERV-L/MERVL Expression by REX1 and Epigenetic Control of Stem Cell Potency. *Frontiers in Oncology*, 4, p.14.

Schuettengruber, B. et al., 2017. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell*, 171(1), pp.34–57.

Schule, K.M. et al., 2019. GADD45 promotes locus-specific DNA demethylation and 2C cycling in embryonic stem cells. *Genes & Development*, 33(13-14), pp.782-798.

Schultz, R.M., 2002. The molecular foundations of the maternal to zygotic transition in the preimplantation embryo. *Human Reproduction Update*, 8(4), pp.323–331.

Schultz, R.M. & Kopf, G.S., 1995. Molecular basis of mammalian egg activation. *Current topics in developmental biology*, 30, pp.21–62.

Sekiguchi, S. et al., 2006. Localization of ubiquitin C-terminal hydrolase L1 in mouse ova and its function in the plasma membrane to block polyspermy. *The American Journal of Pathology*, 169(5), pp.1722–1729.

Sha, Q.-Q. et al., 2018. CNOT6L couples the selective degradation of maternal transcripts to meiotic cell cycle progression in mouse oocyte. *The EMBO journal*, 37(24), p.e99333.

Shalem, O. et al., 2014. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166), pp.84–87.

Sharif, J. et al., 2007. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, 450(7171), pp.908–912.

Sharma, U. et al., 2016. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science*, 351(6271), pp.391–396.

Shen, L. et al., 2014. Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell Stem Cell*, 15(4), pp.459–471.

Sheng, K. et al., 2017. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nature Methods*, 14(3), pp.267–270.

Singh, R. et al., 2015. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Research*, 43(18), p.e118.

Skene, P.J. & Henikoff, S., 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, 6, p.e21856.

Smallwood, S.A. et al., 2011. Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature Genetics*, 43(8), pp.811–814.

Smallwood, S.A. & Kelsey, G., 2012. De novo DNA methylation: a germ cell perspective. *Trends in Genetics*, 28(1), pp.33–42.

Smith, Z.D. et al., 2012. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature*, 484(7394), pp.339–344.

Smith, Z.D. et al., 2014. DNA methylation dynamics of the human preimplantation embryo. *Nature*, 511(7511), pp.611–615.

Sorensen, R.A. & Wassarman, P.M., 1976. Relationship between growth and meiotic maturation of the mouse oocyte. *Developmental Biology*, 50(2), pp.531–536.

Sousa Martins, J.P. et al., 2016. DAZL and CPEB1 regulate mRNA translation synergistically during oocyte maturation. *Journal of Cell Science*, 129(6), pp.1271–1282.

Sripathy, S.P., Stevens, J. & Schultz, D.C., 2006. The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. *Molecular and Cellular Biology*, 26(22), pp.8623–8638.

Stanney, W. et al., 2019. TALE and NF-Y co-occupancy marks enhancers of developmental control genes during zygotic genome activation in zebrafish. *bioRxiv*, p.720102.

Stewart, K.R., Veselovska, L. & Kelsey, G., 2016. Establishment and functions of DNA methylation in the germline. *Epigenomics*, 8(10), pp.1399-1413.

Stoeckius, M. et al., 2018. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome biology*, 19(1), p.224.

Stopka, T. & Skoultchi, A.I., 2011. The ISWI ATPase Snf2h is required for early mouse development. *Proceedings of the National Academy of Sciences*, 100(24), pp.14097–14102.

Storm, M.P. et al., 2014. Zscan4 is regulated by PI3-kinase and DNA-damaging agents and directly interacts with the transcriptional repressors LSD1 and CtBP2 in mouse embryonic stem cells. *PLoS ONE*, 9(3), p.e89821.

Streets, A.M. et al., 2014. Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences*, 111(19), pp.7048–7053.

Su, Y.-Q. et al., 2007. Selective degradation of transcripts during meiotic maturation of mouse oocytes. *Developmental Biology*, 302(1), pp.104–117.

Sun, H. et al., 2018. Methionine adenosyltransferase 2A regulates mouse zygotic genome activation and morula to blastocyst transition. *Biology of Reproduction*, 100(3), pp.601-617.

Sutherland, A.E., Speed, T.P. & Calarco, P.G., 1990. Inner cell allocation in the mouse morula: the role of oriented division during fourth cleavage. *Developmental Biology*, 137(1), pp.13–25.

Sutovsky, P. & Schatten, G., 2000. Paternal contributions to the mammalian zygote: fertilization after sperm-egg fusion. *International review of cytology*, 195, pp.1–65.

Suzumori, N. et al., 2003. RFPL4 interacts with oocyte proteins of the ubiquitin-proteasome degradation pathway. *Proc Natl Acad Sci USA*, 100(2), p.550.

Svensson, V., Vento-Tormo, R. & Teichmann, S.A., 2018. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4), pp.599–604.

Svoboda, P., 2017. Long and small noncoding RNAs during oocyte-to-embryo transition in mammals. *Biochemical Society Transactions*, p.BST20170033.

Svoboda, P. et al., 2004. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Developmental Biology*, 269(1), pp.276–285.

Svoboda, P., Franke, V. & Schultz, R.M., 2015. Chapter Nine - Sculpting the Transcriptome During the Oocyte-to-Embryo Transition in Mouse. In H. D. Lipshitz, ed. *The Maternal-to-Zygotic Transition*. Current Topics in Developmental Biology. Academic Press, pp. 305–349.

Tan, J. & Martin, S.E., 2016. Validation of Synthetic CRISPR Reagents as a Tool for Arrayed Functional Genomic Screening. *PLoS ONE*, 11(12), p.e0168968.

Tanaka, T.S., 2009. Transcriptional heterogeneity in mouse embryonic stem cells. *Reproduction, Fertility and Development*, 21(1), pp.67–75.

Tanenbaum, M.E. et al., 2014. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell*, 159(3), pp.635–646.

Tang, F. et al., 2011. Deterministic and Stochastic Allele Specific Gene Expression in Single Mouse Blastomeres M. Pesce, ed. *PLoS ONE*, 6(6), p.e21208.

Tang, F. et al., 2007. Maternal microRNAs are essential for mouse zygotic development. *Genes & Development*, 21(6), pp.644–648.

Tang, F. et al., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5), pp.377–382.

Tang, F. et al., 2010. RNA-Seq analysis to capture the transcriptome landscape of a single cell. 5(3), pp.516–535.

TARKOWSKI, A.K., 1959. Experiments on the Development of Isolated Blastomeres of Mouse Eggs. *Nature.*, 184(4695), pp.1286–1287.

TARKOWSKI, A.K. et al., 2010. Individual blastomeres of 16- and 32-cell mouse embryos are able to develop into foetuses and mice. *Developmental Biology*, 348(2), pp.190–198.

Tashiro, F. et al., 2010. Maternal-effect gene Ces5/Ooep/Moep19/Floped is essential for oocyte cytoplasmic lattice formation and embryonic development at the maternal-zygotic stage transition. *Genes to cells: devoted to molecular & cellular mechanisms*, 15(8), pp.813–828.

Terns, R.M. & Terns, M.P., 2014. CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends in Genetics*, 30(3), pp.111–118.

Thakore, P.I. et al., 2015. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature Methods*, 12(12), pp.1143–1149.

Thompson, P.J. et al., 2015. hnRNP K Coordinates Transcriptional Silencing by SETDB1 in Embryonic Stem Cells G. S. Barsh, ed. *PLoS Genet.*, 11(1), p.e1004933.

Tong, Z.B. et al., 2000. Mater, a maternal effect gene required for early embryonic development in mice. *Nature Genetics*, 26(3), pp.267–268.

Torres-Padilla, M.-E. & Zernicka-Goetz, M., 2006. Role of TIF1alpha as a modulator of embryonic transcription in the mouse zygote. *J Cell Biol*, 174(3), pp.329–338.

Torres-Padilla, M.-E. et al., 2006. Dynamic distribution of the replacement histone variant H3.3 in the mouse oocyte and preimplantation embryos. *The International journal of developmental biology*, 50(5), pp.455–461.

Torres-Padilla, M.E. & Chambers, I., 2014. Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. *Development*, 141(11), pp.2173–2181.

Töhönen, V. et al., 2015. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nature Communications*, 6, p.8207.

Tsai, T.-C. et al., 2010. Granzyme G is expressed in the two-cell stage mouse embryo and is required for the maternal-zygotic transition. *BMC Developmental Biology*, 10(1), pp.88–16.

Tsukada, Y.-I., Akiyama, T. & Nakayama, K.I., 2015. Maternal TET3 is dispensable for embryonic development but is required for neonatal growth. *Scientific Reports*, 5, p.15876.

TSUKAMOTO, S. et al., 2008. Autophagy Is Essential for Preimplantation Development of Mouse Embryos. *Science*, 321(5885), p.117.

Tumbar, T., Sudlow, G. & Belmont, A.S., 1999. Large-scale chromatin unfolding and remodeling induced by VP16 acidic activation domain. *J Cell Biol*, 145(7), pp.1341–1354.

Uusi-Mäkelä, M.I.E. et al., 2018. Chromatin accessibility is associated with CRISPR-Cas9 efficiency in the zebrafish (Danio rerio) B. B. Riley, ed. *PLoS ONE*, 13(4), p.e0196238.

van de Velde, H. et al., 2008. The four blastomeres of a 4-cell stage human embryo are able to develop individually into blastocysts with inner cell mass and trophectoderm. *Human reproduction (Oxford, England)*, 23(8), pp.1742–1747.

van der Heijden, G.W. et al., 2005. Asymmetry in histone H3 variants and lysine methylation between paternal and maternal chromatin of the early mouse zygote. *Mechanisms of Development*, 122(9), pp.1008–1022.

van Essen, D. et al., 2009. Two modes of transcriptional activation at native promoters by NF-kappaB p65. *PLoS Biol*, 7(3), p.e73.

van Steensel, B., Delrow, J. & Henikoff, S., 2001. Chromatin profiling using targeted DNA adenine methyltransferase. *Nature Genetics*, 27(3), pp.304–308.

Vassena, R. et al., 2011. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development*, 138(17), pp.3699–3709.

Vasseur, M., Condamine, H. & Duprey, P., 1985. RNAs containing B2 repeated sequences are transcribed in the early stages of mouse embryogenesis. *The EMBO journal*, 4(7), pp.1749–1753.

Vastenhouw, N.L., Cao, W.X. & Lipshitz, H.D., 2019. The maternal-to-zygotic transition revisited. *Development*, 146(11), p.dev161471.

Venkatesh, S. & Workman, J.L., 2015. Histone exchange, chromatin structure and the regulation of transcription. *Nature Reviews Molecular Cell Biology*, 16(3), pp.178–189.

Vento-Tormo, R. et al., 2018. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature*, 563(7731), pp.347–353.

Verboom, K. et al., 2019. SMARTer single cell total RNA sequencing. *Nucleic Acids Research*, pii:gkz535.

Veselovska, L. et al., 2015. Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. *Genome biology*, 16(1), pp.20110330–0–17.

Vidigal, J.A. & Ventura, A., 2015. Rapid and efficient one-step generation of paired gRNA CRISPR-Cas9 libraries. *Nature Communications*, 6, p.8083.

Vignali, M. et al., 2000. ATP-dependent chromatin-remodeling complexes. *Molecular and Cellular Biology*, 20(6), pp.1899–1910.

Vitak, S.A. et al., 2017. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, 14(3), pp.302–308.

Vitale, A.M. et al., 2007. Proteomic profiling of murine oocyte maturation. *Mol Reprod Dev*, 74(5), pp.608–616.

Vojta, A. et al., 2016. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Research*, 44(12), pp.5615–5628.

Vuoristo, S. et al., 2019. DUX4 regulates oocyte to embryo transition in human. *bioRxiv*, p.732289.

Walser, C.B. & Lipshitz, H.D., 2011. Transcript clearance during the maternal-to-zygotic transition. *Current Opinion in Genetics & Development*, 21(4), pp.431–443.

Wan, L.-B., Pan, H., Hannenhalli, S., Cheng, Y., Ma, J., Fedoriw, A., Lobanenkov, V., Latham, K.E., Schultz, R.M. & Bartolomei, M.S., 2008. Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. *Development*, 135(16), pp.2729–2738.

Wang, B. et al., 2016. Proteomic Analysis of Mouse Oocytes Identifies PRMT7 as a Reprogramming Factor that Replaces SOX2 in the Induction of Pluripotent Stem Cells. *Journal of Proteome Research*, 15(8), pp.2407–2421.

Wang, Jiaqiang et al., 2016. A novel long intergenic noncoding RNA indispensable for the cleavage of mouse two-cell embryos. *EMBO reports*, p.e201642051.

Wang, Jichang et al., 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, 516(7531), pp.405–409.

Wang, L. et al., 2014. Programming and Inheritance of Parental DNA Methylomes in Mammals. *Cell.*, 157(4), pp.979–991.

Wang, T. et al., 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, 343(6166), pp.80–84.

Wassarman, P.M. & Kinloch, R.A., 1992. Gene expression during oogenesis in mice. *Mutation research*, 296(1-2), pp.3–15.

Wasson, J.A. et al., 2016. Maternally provided LSD1/KDM1A enables the maternal-to-zygotic transition and prevents defects that manifest postnatally. *eLife*, 5.

Waterston, R.H. et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–562.

Wei, Y., Terns, R.M. & Terns, M.P., 2015. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes & Development*, 29(4), pp.356–361.

Weltner, J. et al., 2018. Human pluripotent reprogramming with CRISPR activators. *Nature Communications*, 9(1), p.2643.

Wennekamp, S. et al., 2013. A self-organization framework for symmetry breaking in the mammalian embryo. *Nature Reviews Molecular Cell Biology*, 14(7), pp.454–461.

Whiddon, J.L. et al., 2017. Conservation and innovation in the DUX4-family gene network. *Nature Genetics*, 49(6), pp.935-940.

Wicker, T. et al., 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12), pp.973–982.

Wickramasinghe, D., Ebert, K.M. & Albertini, D.F., 1991. Meiotic competence acquisition is associated with the appearance of M-phase characteristics in growing mouse oocytes. *Developmental Biology*, 143(1), pp.162–172.

Wiekowski, M., Miranda, M. & DePamphilis, M.L., 1991. Regulation of gene expression in preimplantation mouse embryos: effects of the zygotic clock and the first mitosis on promoter and enhancer activities. *Developmental Biology*, 147(2), pp.403–414.

Wiekowski, M., Miranda, M. & DePamphilis, M.L., 1993. Requirements for promoter activity in mouse oocytes and embryos distinguish paternal pronuclei from maternal and zygotic nuclei. *Developmental Biology*, 159(1), pp.366–378.

Wolf, F.A., Angerer, P. & Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1), p.15.

Wong, A.K. & Rattner, J.B., 1988. Sequence organization and cytological localization of the minor satellite of mouse. *Nucleic Acids Research*, 16(24), pp.11645–11661.

Wongtawan, T. et al., 2011. Histone H4K20me3 and HP1~ are late heterochromatin markers in development, but present in undifferentiated embryonic stem cells. *Journal of Cell Science*, 124(11), pp.1878–1890.

Wossidlo, M. et al., 2011. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature Communications*, 2, p.241.

Wossidlo, M. et al., 2010. Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *The EMBO journal*, 29(11), pp.1877–1888.

Wu, H. & Zhang, Yi, 2014. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, 156(1-2), pp.45–68.

Wu, J. et al., 2018. Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature*, 557(7704), pp.256–260.

Wu, J. et al., 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*, 534(7609), pp.652–657.

Wu, Xuebing et al., 2014. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnology*, 32, pp.670-676.

Wu, Xuemei et al., 2003. Zygote arrest 1 (Zar1) is a novel maternal-effect gene critical for the oocyte-to-embryo transition. *Nature Genetics*, 33(2), pp.187–191.

Xia, W. et al., 2019. Resetting histone modifications during human parental-to-zygotic transition. *Science*, 365(6451), p.353.

Xie, L. et al., 2011. KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. *The EMBO journal*, 30(8), pp.1473–1484.

Xie, S. et al., 2018. Frequent sgRNA-barcode recombination in single-cell perturbation assays W. Hu, ed. *PLoS ONE*, 13(6), p.e0198635.

Xie, S. et al., 2017. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell*, 66(2), pp.285–299.e5.

Xu, H. et al., 2015. Sequence determinants of improved CRISPR sgRNA design. *Genome Research*, 25(8), pp.1147–1157.

Xu, Q. et al., 2015. Maternal BCAS2 protects genomic integrity in mouse early embryonic development. *Development*, 142(22), pp.3943–3953.

Xu, Q. et al., 2019. SETD2 regulates the maternal epigenome, genomic imprinting and embryonic development. *Nature Genetics*.

Xu, X., Duan, D. & Chen, S.-J., 2017. CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: from physical mechanism to off-target assessment. *Scientific Reports*, 7(1), pp.143–143.

Xu, Y.-W. et al., 2017. Maternal DCAF2 is crucial for maintenance of genome stability during the first cell cycle in mice. *Journal of Cell Science*, 130(19), pp.3297-3307.

Xue, Z. et al., 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature.*, 500(7464), pp.593–597.

Yan, L. et al., 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structure Molecular Biology*, 20(9), pp.1131–1139.

Yan, Y.-L. et al., 2019. DPPA2/4 and SUMO E3 ligase PIAS4 opposingly regulate zygotic transcriptional program. *PLoS Biol*, 17(6), p.e3000324.

Yang, J. et al., 2019. Genome-Scale CRISPRa Screen Identifies Novel Factors for Cellular Reprogramming. *Stem Cell Reports*, 12(4), pp.757–771.

Yang, L. et al., 2018. KDM6A and KDM6B play contrasting roles in nuclear transfer embryos revealed by MERVL reporter system. *EMBO reports*, 19(12), pii:e46240.

Yang, L. et al., 2016. The Maternal Effect Genes UTX and JMJD3 Play Contrasting Roles in Mus musculus Preimplantation Embryo Development. *Scientific Reports*, 6, p.26711.

Yang, Y. et al., 2017. The E3 ubiquitin ligase RNF114 and TAB1 degradation are required for maternal-to-zygotic transition. *EMBO reports*, p.e201642573.

Yao, J. & Dai, H.-L., 2018. Is Pooled CRISPR-Screening the Dawn of a New Era for Functional Genomics. *Advances in experimental medicine and biology*, 1068, pp.171–176.

Yao, S. et al., 2004. Retrovirus Silencing, Variegation, Extinction, and Memory Are Controlled by a Dynamic Interplay of Multiple Epigenetic Modifications. *Molecular Therapy*, 10(1), pp.27–36.

Yarrington, R.M. et al., 2018. Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proc Natl Acad Sci USA*, 115(38), p.9351.

Yartseva, V. & Giraldez, A.J., 2015. The Maternal-to-Zygotic Transition During Vertebrate Development. In *The Maternal-to-Zygotic Transition*. *Current Topics in Developmental Biology*. *Elsevier,* pp. 191–232.

Yu, C. et al., 2013. CRL4 complex regulates mammalian oocyte survival and reprogramming by activation of TET proteins. *Science*, 342(6165), pp.1518–1521.

Yu, C., Ji, S.-Y., Dang, Y.-J., et al., 2016. Oocyte-expressed yes-associated protein is a key activator of the early zygotic genome in mouse. *Cell Research*, 26(3), pp.275–287.

Yu, C., Ji, S.-Y., Sha, Q.-Q., et al., 2016. BTG4 is a meiotic cell cycle-coupled maternal-zygotic-transition licensing factor in oocytes. *Nature Structural & Molecular Biology*, 23(5), pp.387–394.

Yu, J., Hecht, N.B. & Schultz, R.M., 2001. Expression of MSY2 in Mouse Oocytes and Preimplantation Embryos1. *Biology of Reproduction*, 65(4), pp.1260–1270.

Yu, J., Hecht, N.B. & Schultz, R.M., 2002. RNA-Binding Properties and Translation Repression In Vitro by Germ Cell-Specific MSY2 Protein1. *Biology of Reproduction*, 67(4), pp.1093–1098.

Yu, X.-J. et al., 2014. The subcortical maternal complex controls symmetric division of mouse zygotes by regulating F-actin dynamics. *Nature Communications*, 5, p.4887.

Yuen, G. et al., 2017. CRISPR/Cas9-mediated gene knockout is insensitive to target copy number but is dependent on guide RNA potency and Cas9/sgRNA threshold expression level. *Nucleic Acids Research*, 45(20), pp.12039–12053.

Yurttas, P. et al., 2008. Role for PADI6 and the cytoplasmic lattices in ribosomal storage in oocytes and translational control in the early mouse embryo. *Development*, 135(15), pp.2627–2636.

Zalatan, J.G. et al., 2015. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell*, 160(1-2), pp.339–350.

Zalzman, M. et al., 2010. Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature*, 464(7290), pp.858–863.

Zamboni, L., Chakraborty, J. & Smith, D.M., 1972. First Cleavage Division of the Mouse Zygote: An Ultrastructural Study1. *Biology of Reproduction*, 7(2), pp.170–193.

Zeng, F. & Schultz, R.M., 2005. RNA transcript profiling during zygotic gene activation in the preimplantation mouse embryo. *Developmental Biology*, 283(1), pp.40–57.

Zeng, F., Baldwin, D.A. & Schultz, R.M., 2004. Transcript profiling during preimplantation mouse development. *Developmental Biology*, 272(2), pp.483–496.

Zeng, T.-B. et al., 2019. EHMT2 and SETDB1 protect the maternal pronucleus from 5mC oxidation. *Proceedings of the National Academy of Sciences*, 116(22), pp.10834-10841.

Zhang, B. et al., 2016. Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature*, 537(7621), pp.553–557.

Zhang, C. et al., 2007. GATA factors induce mouse embryonic stem cell differentiation toward extraembryonic endoderm. *Stem cells and development*, 16(4), pp.605–613.

Zhang, K. et al., 2014. Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. *BMC Genomics*, 15(1), p.845.

Zhang, P. et al., 2009. Transcriptome profiling of human pre-implantation development. *PLoS ONE*, 4(11), p.e7844.

Zhang, W. et al., 2019. Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. *Nucleic Acids Research*, pii:gkz594.

Zhang, X. et al., 2019. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, 73(1), pp.130–142.e5.

Zhang, Yanxiao et al., 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics*, 51(9), pp.1380-1388.

Zhao, T. et al., 2018. Single-Cell RNA-Seq Reveals Dynamic Early Embryonic-like Programs during Chemical Reprogramming. *Cell Stem Cell*, 23(1), pp.31–45.e7.

Zheng, G.X.Y. et al., 2017. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8, p.14049.

Zheng, H. et al., 2016. Resetting Epigenetic Memory by Reprogramming of Histone Modifications in Mammals. *Molecular Cell*, 63(6), pp.1066–1079.

Zheng, P. & Dean, J., 2009. Role of Filia, a maternal effect gene, in maintaining euploidy during cleavage-stage mouse embryogenesis. *Proceedings of the National Academy of Sciences*, 106(18), pp.7473–7478.

Zhou, L.-Q. & Dean, J., 2015. Reprogramming the genome to totipotency in mouse embryos. *Trends in Cell Biology*, 25(2), pp.82–91.

Zhou, V.W., Goren, A. & Bernstein, B.E., 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 12(1), pp.7–18.

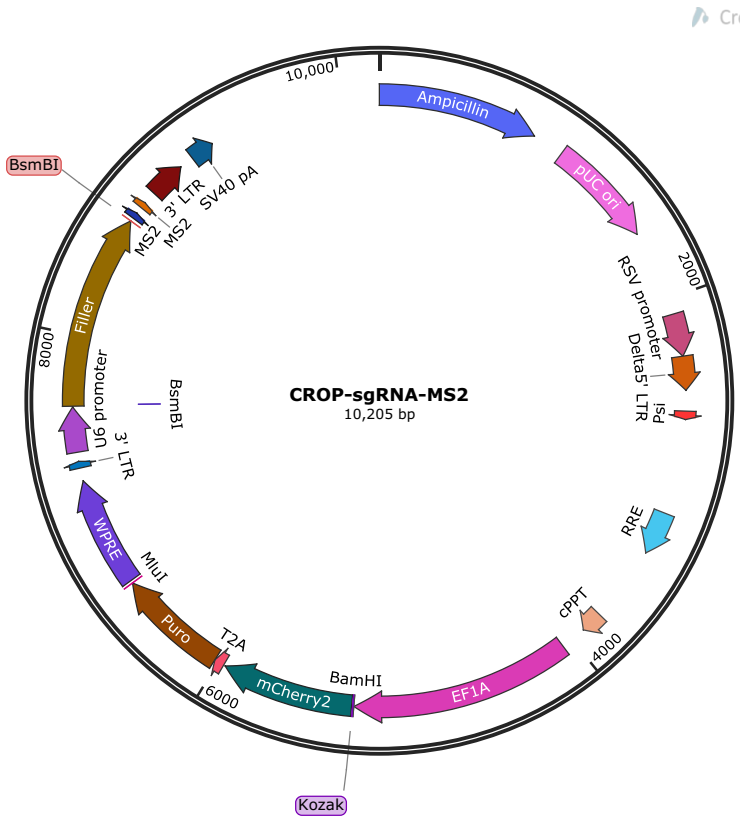
Zhu, C. et al., 2017. Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell*, 20(5), pp.720-731.

Ziegenhain, C. et al., 2017. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65(4), pp.631–643.e4.

Zilionis, R. et al., 2017. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1), pp.44–73.

Zuccotti, M. et al., 1995. Chromatin organization during mouse oocyte growth. *Mol Reprod Dev*, 41(4), pp.479–485.

# Appendix A CROP-sgRNA-MS2 plasmid sequence



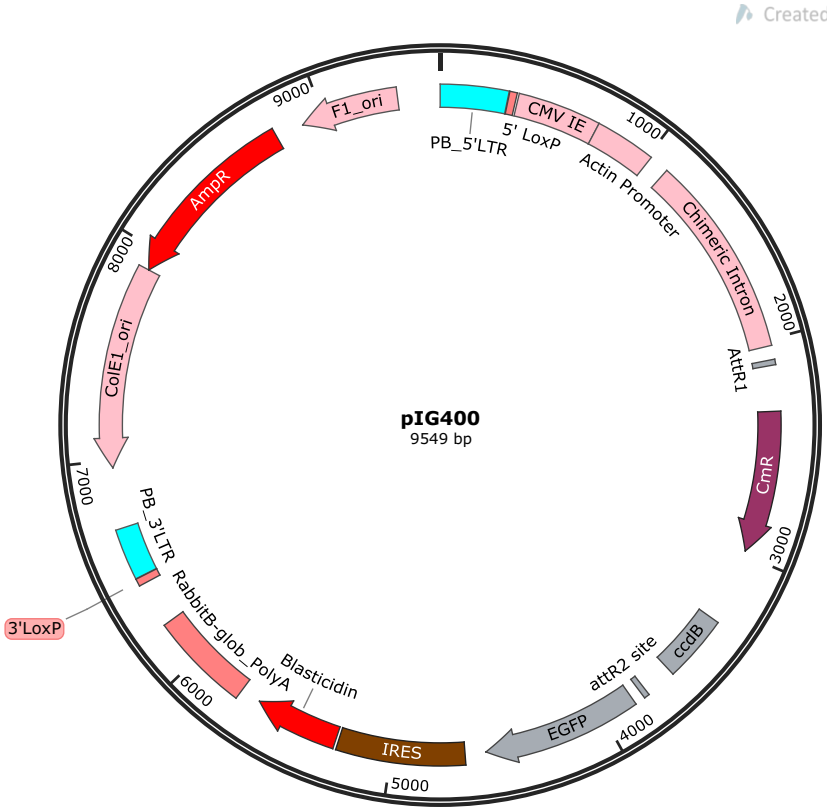

Map of CROP-sgRNA-MS2 vector

5’atgagtattcaacatttccgtgtcgcccttattccctttttgcggcattttgccttcctgtttttgctcacccagaaacgctggtgaaagtaaaagatgc
tgaagatcagttgggtgcacgagtgggttacatcgaactggatctcaacagcggtaagatccttgagagttttcgccccgaagaacgttttccaatga
tgagcacttttaaagttctgctatgtggcgcggtattatcccgtattgacgccgggcaagagcaactcggtcgccgcatacactattctcagaatgact
tggttgagtactcaccagtcacagaaaagcatcttacggatggcatgacagtaagagaattatgcagtgctgccataaccatgagtgataacactgc
ggccaacttacttctgacaacgatcggaggaccgaaggagctaaccgctttttgcacaacatgggggatcatgtaactcgccttgatcgttgggaa
ccggagctgaatgaagccataccaaacgacgagcgtgacaccacgatgcctgtagcaatggcaacaacgttgcgcaaactattaactggcgaact
acttactctagcttcccggcaacaattaatagactggatggaggcggataaagttgcaggaccacttctgcgctcggcccttccggctggctggttta
ttgctgataaatctggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccctcccgtatcgtagttatctacac
gacggggagtcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgattaagcattggtaactgtcagaccaagttta
ctcatatatactttagattgatttaaaacttcatttttaatttaaaaggatctaggtgaagatccttttgataatctcatgaccaaaatcccttaacgtgagttt
tcgttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatccttttttttctgcgcgtaatctgctgcttgcaaacaaaaaaaccac
cgctaccagcggtggtttgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgttcttc
tagtgtagccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgttaccagtggctgctgccagtggc

gataagtcgtgtcttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacgggggttcgtgcacacagccca
gcttggagcgaacgacctacaccgaactgagatacctacagcgtgagctatgagaaagcgccacgcttcccgaagggagaaaggcggacaggt
atccggtaagcggcagggtcggaacaggagagcgcacgagggagcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgcca
cctctgacttgagcgtcgatttttgtgatgctcgtcaggggggcggagcctatggaaaaacgccagcaacgcggccttttacggttcctggccttttg
ctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataaccgtattaccgcctttgagtgagctgataccgctcgccgcagccgaa
cgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcccaatacgcaaaccgcctctccccgcgcgttggccgattcattaatgca
gctggcacgacaggtttcccgactggaaagcgggcagtgagcgcaacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactt
tatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcacacaggaaacagctatgaccatgattacgccaagcgcgcaattaacc
ctcactaaagggaacaaaagctggagctgcaagcttaatgtagtcttatgcaatactcttgtagtcttgcaacatggtaacgatgagttagcaacatgc
cttacaaggagagaaaaagcaccgtgcatgccgattggtggaagtaaggtggtacgatcgtgccttattaggaaggcaacagacgggtctgacat
ggattggacgaaccactgaattgccgcattgcagagatattgtatttaagtgcctagctcgatacataaacgggtctctctggttagaccagatctgag
cctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgcttcaagtagtgtgtgcccgtctgttgtgtgact
ctggtaactagagatccctcagacccttttagtcagtgtggaaaatctctagcagtggcgcccgaacagggacttgaaagcgaaagggaaaccag
aggagctctctcgacgcaggactcggcttgctgaagcgcgcacggcaagaggcgaggggcggcgactggtgagtacgccaaaaattttgacta
gcggaggctagaaggagagagatgggtgcgagagcgtcagtattaagcgggggagaattagatcgcgatgggaaaaaattcggttaaggccag
ggggaaagaaaaaatataaattaaaacatatagtatgggcaagcagggagctagaacgattcgcagttaatcctggcctgttagaaacatcagaag
gctgtagacaaatactgggacagctacaaccatcccttcagacaggatcagaagaacttagatcattatataatacagtagcaaccctctattgtgtgc
atcaaaggatagagataaaagacaccaaggaagctttagacaagatagaggaagagcaaaacaaaagtaagaccaccgcacagcaagcggcc
gctgatcttcagacctggaggaggagatatgagggacaattggagaagtgaattatataaatataaagtagtaaaaattgaaccattaggagtagca
cccaccaaggcaaagagaagagtggtgcagagagaaaaaagagcagtgggaataggagctttgttccttgggttcttgggagcagcaggaagc
actatgggcgcagcgtcaatgacgctgacggtacaggccagacaattattgtctggtatagtgcagcagcagaacaatttgctgagggctattgag
gcgcaacagcatctgttgcaactcacagtctggggcatcaagcagctccaggcaagaatcctggctgtggaaagatacctaaaggatcaacagct
cctggggatttggggttgctctggaaaactcatttgcaccactgctgtgccttggaatgctagttggagtaataaatctctggaacagatttggaatcac
acgacctggatggagtgggacagagaaattaacaattacacaagcttaatacactccttaattgaagaatcgcaaaaccagcaagaaaagaatgaa
caagaattattggaattagataaatgggcaagtttgtggaattggtttaacataacaaattggctgtggtatataaaattattcataatgatagtaggagg
cttggtaggtttaagaatagtttttgctgtactttctatagtgaatagagttaggcagggatattcaccattatcgtttcagaatcttgagacaaatggcagt
attcatccacaattttaaaagaaaaggggggattgggggtacagtgcaggggaaagaatagtagacataatagcaacagacatacaaactaaag
aattacaaaaacaaattacaaaaattcaaaattttcgggtttattacagggacagcagagatccactttggcgccggctcgaggggcccgggtgca
aagatggataaagttttaaacagagaggaatctttgcagctaatggaccttctaggtcttgaaaggagtgggaattggctccggtgcccgtcagtgg
gcagagcgcacatcgcccacagtccccgagaagttgggggggaggggtcggcaattgatccggtgcctagagaaggtggcgcggggtaaactg
ggaaagtgatgtcgtgtactggctccgccttttcccgagggtgggggagaaccgtatataagtgcagtagtcgccgtgaacgttcttttcgcaacg
ggtttgccgccagaacacaggtaagtgccgtgtgtggttcccgcgggcctggcctctttacgggttatggcccttgcgtgccttgaattacttccacct
ggctgcagtacgtgattcttgatcccgagcttcgggttggaagtgggtgggagagttcgaggccttgcgcttaaggagccccttcgcctcgtgcttg
agttgaggcctggcctgggcgctggggccgccgcgtgcgaatctggtggcaccttcgcgcctgtctcgctgctttcgataagtctctagccatttaa
aatttttgatgacctgctgcgacgctttttttctggcaagatagtcttgtaaatgcgggccaagatctgcacactggtatttcggttttggggccgcggg
cggcgacggggcccgtgcgtcccagcgcacatgttcggcgaggcggggcctgcgagcgcggccaccgagaatcggacgggggtagtctcaa
gctggccggcctgctctggtgcctggcctcgcgccgccgtgtatcgccccgccctgggcggcaaggctggcccggtcggcaccagttgcgtga
gcggaaagatggccgcttcccggccctgctgcagggagctcaaaatggaggacgcggcgctcgggagagcgggcgggtgagtcacccacac
aaaggaaaagggcctttccgtcctcagccgtcgcttcatgtgactccacggagtaccgggcgccgtccaggcacctcgattagttctcgagcttttg
gagtacgtcgtctttaggttgggggagggtttttatgcgatggagtttccccacactgagtgggtggagactgaagttaggccagcttggcacttg
atgtaattctccttggaatttgccctttttgagtttggatcttggttcattctcaagcctcagacagtggttcaaagttttttttcttccatttcaggtgtcgtgag
gatccgccaccatggtgagcaagggcgaggaggataacatggccatcatcaaggagttcatgcgcttcaaggtgcacatggagggctccgtgaa
cggccacgagttcgagatcgagggcgagggcgagggccgcccctacgagggcacccagaccgccaagctgaaggtgaccaagggcggccc
cctgcccttcgcctgggacatcctgtcccctcagttcatgtacggctccaaggcctacgtgaagcaccccgccgacatccccgactacctgaagct
gtccttccccgagggcttcaattgggagcgcgtgatgaacttcgaggacggcggcgtggtgaccgtgacccaggactcctccctgcaggacggc
gagttcatctacaaggtgaagctgcgcggcaccaacttcccctccgacggccccgtgatgcagtgtaggaccatgggctgggaggcctccactga
gcggatgtaccccgaggacggcgccctgaagggcgagatcaagcagaggctgaagctgaaggacggcggccactacgacgctgaggtgaag
accacctacaaggccaagaagcccgtgcagctgcccggcgcctacaacgtggacatcaagctggacatcctgtcccacaacgaggactacacc
atcgtggagcagtacgagcgcgccgagggccgccactccaccggcggcatggacgagctgtacaagggcagcggagaaggtagaggttctct
cctcacttgtggtgatgttgaagaaaaccctggtccaatgaccgagtacaagcccacggtgcgcctcgccacccgcgacgacgtccccagggcc

gtacgcaccctcgccgccgcgttcgccgactaccccgccacgcgccacaccgtcgatccggaccgccacatcgagcgggtcaccgagctgcaa
gaactcttcctcacgcgcgtcgggctcgacatcggcaaggtgtgggtcgcggacgacggcgccgccgtggcggtctggaccacgccggagag
cgtcgaagcgggggcggtgttcgccgagatcggcccgcgcatggccgagttgagcggttcccggctggccgcgcagcaacagatggaaggcc
tcctggcgccgcaccggcccaaggagcccgcgtggttcctggccaccgtcggagtctcgcccgaccaccagggcaagggtctgggcagcgcc
gtcgtgctccccggagtggaggcggccgagcgcgccgggggtgcccgccttcctggagacctccgcgccccgcaacctccccttctacgagcgg
ctcggcttcaccgtcaccgccgacgtcgaggtgcccgaaggaccgcgcacctggtgcatgacccgcaagcccggtgcctgaacgcgttaagtcg
acaatcaacctctggattacaaaatttgtgaaagattgactggtattcttaactatgttgctccttttacgctatgtggatacgctgctttaatgcctttgtatc
atgctattgcttcccgtatggctttcattttctcctccttgtataaatcctggttgctgtctctttatgaggagttgtggcccgttgtcaggcaacgtggcgtg
gtgtgcactgtgtttgctgacgcaacccccactggttggggcattgccaccacctgtcagctcctttccgggactttcgctttccccctccctattgcca
cggcggaactcatcgccgcctgccttgcccgctgctggacaggggctcggctgttgggcactgacaattccgtggtgttgtcggggaaatcatcgt
cctttccttggctgctcgcctgtgttgccacctggattctgcgcgggacgtccttctgctacgtcccttcggccctcaatccagcggaccttccttcccg
cggcctgctgccggctctgcggcctcttccgcgtcttcgccttcgccctcagacgagtcggatctccctttgggccgcctccccgcgtcgacttttaag
accaatgacttacaaggcagctgtagatcttagccacttttaaaagaaaaggggggactggaagggctaattcactcccaacgaagacaagatttc
accattatcgtttcagacccacctcccaaccccgaggggacccagagagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgt
tagagagataattagaattaatttgactgtaaacacaaagatattagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagttttaaa
attatgttttaaaatggactatcatatgcttaccgtaacttgaaagtatttcgatttcttggctttatatatcttgtggaaaggacgaaacaccggagacggt
tgtaaatgagcacacaaaatacacatgctaaaatattatattctatgacctttataaaatcaaccaaaatcttctttttaataactttagtatcaataattagaa
ttttatgttccttttgcaaacttttaataaaaatgagcaaaataaaaaaacgctagttttagtaactcgcgttgttttcttcacctttaataatagctactcca
ccacttgttcctaagcggtcagctcctgcttcaatcatttttgagcatcttcaaatgttctaactccaccagctgctttaactaaagcattgtctttaacaac
tgacttcattagtttaacatcttcaaatgttgcacctgattttgaaaatcctgttgatgttttaacaaattctaatccagcttcaacagctatttcacaagctttc
atgatttcttctttgttaataaacaattttccataatacatttaacaacatgtgatccagctgctttttttacagctttcatgtcttctaaaactaattcataatttt
gtcttttaatgcaccaatatttaataccatatcaatttctgttgcaccatctttaattgcttcagaaacttcgaatgcttttgtagctgttgtgcatgcacctag
aggaaaacctacaacatttgttattcctacatttgtgccttttaataattctttacaatagcttgttcaatatgaattaacacaaactgttgcaaaatcaaattc
aattgcttcatcacataattgtttaatttcagctttcgtagcatcttgttttaataatgtgtgatctatatatttgtttagtttcatttttttctcctatatattcatttttaa
ttttaattctttaataatttcgtctactttaactttagcgttttgaacagattcaccaacacctataaaataaatttttagtttaggttcagttccacttgggcgaa
cagcaaatcatgacttatcttctaaataaaatttttagtaagtcttgtcctggcatattatacattccatcgacgtctctgtttaagagctaggccaacatgag
gatcacccatgtctgcagggcctagcaagtttaaataaggctagtccgttatcaacttggccaacatgaggatcacccatgtctgcagggccaagtg
gcaccgagtcggtgctttttaagcttggcgcaccaagcttggcgtaactagatcttgagacactgctttttgcttgtactgggtctctctggttagacca
gatctgagcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgcttcaagtagtgtgtgcccgtctgtt
gtgtgactctggtaactagagatccctcagacccttttagtcagtgtggaaaatctctagcagtacgtatagtagttcatgtcatcttattattcagtatttat
aacttgcaaagaaatgaatatcagagagtgagaggaacttgtttattgcagcttataatggttacaaataaagcaatagcatcacaaatttcacaaataa
agcatttttttcactgcattctagttgtggtttgtccaaactcatcaatgtatcttatcatgtctggctctagctatcccgcccctaactccgcccgttccgcc
cattctccgccccatggctgactaatttttttttatttatgcagaggccgaggccgcctcggcctctgagctattccagaagtagtgaggaggctttttgg
aggcctagggacgtacccaattcgccctatagtgagtcgtattacgcgcgctcactggccgtcgttttacaacgtcgtgactgggaaaaccctggcg
ttacccaacttaatcgccttgcagcacatccccctttcgccagctggcgtaatagcgaagaggcccgcaccgatcgcccttcccaacagttgcgcag
cctgaatggcgaatgggacgcgccctgtagcggcgcattaagcgcggcgggtgtggtggttacgcgcagcgtgaccgctacacttgccagcgc
cctagcgcccgctcctttcgctttcttcccttcctttctcgccacgttcgccggctttccccgtcaagctctaaatcgggggctccctttagggttccgatt
tagtgctttacggcacctcgaccccaaaaaacttgattagggtgatggttcacgtagtgggccatcgccctgatagacggtttttcgccctttgacgtt
ggagtccacgttctttaatagtggactcttgttccaaactggaacaacactcaaccctatctcggtctattcttttgatttataagggattttgccgatttcg
gcctattggttaaaaaatgagctgatttaacaaaaatttaacgcgaattttaacaaaatattaacgcttacaatttaggtggcacttttcggggaaatgtgc
gcggaacccctatttgtttatttttctaaatacattcaaatatgtatccgctcatgagacaataaccctgataaatgcttcaataatattgaaaaaggaaga
gt3’

# Appendix B pIG400 plasmid sequence




Map of pIG400 vector

5’atgagtattcaacatttccgtgtcgcccttattccctttttgcggcattttgccttcctgtttttgctcacccagaaacgctggtgaaagtaaaagatgc
tgaagatcagttgggtgcacgagtgggttacatcgaactggatctcaacagcggtaagatccttgagagttttcgccccgaagaacgttttccaatga
tgagcacttttaaagttctgctatgtggcgcggtattatcccgtattgacgccgggcaagagcaactcggtcgccgcatacactattctcagaatgact
tggttgagtactcaccagtcacagaaaagcatcttacggatggcatgacagtaagagaattatgcagtgctgccataaccatgagtgataacactgc
ggccaacttacttctgacaacgatcggaggaccgaaggagctaaccgcttttttgcacaacatgggggatcatgtaactcgccttgatcgttgggaa
ccggagctgaatgaagccataccaaacgacgagcgtgacaccacgatgcctgtagcaatggcaacaacgttgcgcaaactattaactggcgaact
acttactctagcttcccggcaacaattaatagactggatggaggcggataaagttgcaggaccacttctgcgctcggcccttccggctggctggttta
ttgctgataaatctggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccctcccgtatcgtagttatctacac
gacggggagtcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgattaagcattggtaactgtcagaccaagttta
ctcatatatactttagattgatttaaaacttcatttttaatttaaaaggatctaggtgaagatccttttgataatctcatgaccaaaatcccttaacgtgagttt
tcgttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatccttttttctgcgcgtaatctgctgcttgcaaacaaaaaaaccac
cgctaccagcggtggtttgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgttcttc
tagtgtagccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgttaccagtggctgctgccagtggc

gataagtcgtgtcttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacgggggttcgtgcacacagccca
gcttggagcgaacgacctacaccgaactgagatacctacagcgtgagctatgagaaagcgccacgcttcccgaagggagaaaggcggacaggt
atccggtaagcggcagggtcggaacaggagagcgcacgagggagcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgcca
cctctgacttgagcgtcgatttttgtgatgctcgtcaggggggcggagcctatggaaaaacgccagcaacgcggccttttacggttcctggccttttg
ctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataaccgtattaccgcctttgagtgagctgataccgctcgccgcagccgaa
cgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcccaatacgcaaaccgcctctccccgcgcgttggccgattcattaatgca
gctggcacgacaggtttcccgactggaaagcgggcagtgagcgcaacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactt
tatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcacacaggaaacagctatgaccatgattacgccaagcgcgcaattaacc
ctcactaaagggaacaaaagctggagctgcaagcttaatgtagtcttatgcaatactcttgtagtcttgcaacatggtaacgatgagttagcaacatgc
cttacaaggagagaaaaagcaccgtgcatgccgattggtggaagtaaggtggtacgatcgtgccttattaggaaggcaacagacgggtctgacat
ggattggacgaaccactgaattgccgcattgcagagatattgtatttaagtgcctagctcgatacataaacgggtctctctggttagaccagatctgag
cctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgcttcaagtagtgtgtgcccgtctgttgtgtgact
ctggtaactagagatccctcagacccttttagtcagtgtggaaaatctctagcagtggcgcccgaacagggacttgaaagcgaaagggaaaccag
aggagctctctcgacgcaggactcggcttgctgaagcgcgcacggcaagaggcgaggggcggcgactggtgagtacgccaaaaattttgacta
gcggaggctagaaggagagagatgggtgcgagagcgtcagtattaagcgggggagaattagatcgcgatgggaaaaaattcggttaaggccag
ggggaaagaaaaaatataaattaaaacatatagtatgggcaagcagggagctagaacgattcgcagttaatcctggcctgttagaaacatcagaag
gctgtagacaaatactgggacagctacaaccatcccttcagacaggatcagaagaacttagatcattatataatacagtagcaaccctctattgtgtgc
atcaaaggatagagataaaagacaccaaggaagctttagacaagatagaggaagagcaaaacaaaagtaagaccaccgcacagcaagcggcc
gctgatcttcagacctggaggaggagatatgagggacaattggagaagtgaattatataaatataaagtagtaaaaattgaaccattaggagtagca
cccaccaaggcaaagagaagagtggtgcagagagaaaaaagagcagtgggaataggagctttgttccttgggttcttgggagcagcaggaagc
actatgggcgcagcgtcaatgacgctgacggtacaggccagacaattattgtctggtatagtgcagcagcagaacaatttgctgagggctattgag
gcgcaacagcatctgttgcaactcacagtctggggcatcaagcagctccaggcaagaatcctggctgtggaaagatacctaaaggatcaacagct
cctggggatttggggttgctctggaaaactcatttgcaccactgctgtgccttggaatgctagttggagtaataaatctctggaacagatttggaatcac
acgacctggatggagtgggacagagaaattaacaattacacaagcttaatacactccttaattgaagaatcgcaaaaccagcaagaaaagaatgaa
caagaattattggaattagataaatgggcaagtttgtggaattggtttaacataacaaattggctgtggtatataaaattattcataatgatagtaggagg
cttggtaggtttaagaatagtttttgctgtactttctatagtgaatagagttaggcagggatattcaccattatcgtttcagaatcttgagacaaatggcagt
attcatccacaattttaaaagaaaaggggggattgggggtacagtgcaggggaaagaatagtagacataatagcaacagacatacaaactaaag
aattacaaaaacaaattacaaaaattcaaaattttcgggtttattacagggacagcagagatccactttggcgccggctcgaggggcccgggtgca
aagatggataaagttttaaacagagaggaatctttgcagctaatggaccttctaggtcttgaaaggagtgggaattggctccggtgcccgtcagtgg
gcagagcgcacatcgcccacagtccccgagaagttgggggggaggggtcggcaattgatccggtgcctagagaaggtggcgcggggtaaactg
ggaaagtgatgtcgtgtactggctccgccttttcccgagggtgggggagaaccgtatataagtgcagtagtcgccgtgaacgttcttttcgcaacg
ggtttgccgccagaacacaggtaagtgccgtgtgtggttcccgcgggcctggcctctttacgggttatggcccttgcgtgccttgaattacttccacct
ggctgcagtacgtgattcttgatcccgagcttcgggttggaagtgggtgggagagttcgaggccttgcgcttaaggagccccttcgcctcgtgcttg
agttgaggcctggcctgggcgctggggccgccgcgtgcgaatctggtggcaccttcgcgcctgtctcgctgctttcgataagtctctagccatttaa
aatttttgatgacctgctgcgacgctttttttctggcaagatagtcttgtaaatgcgggccaagatctgcacactggtatttcggttttttggggccgcggg
cggcgacggggcccgtgcgtcccagcgcacatgttcggcgaggcggggcctgcgagcgcggccaccgagaatcggacgggggtagtctcaa
gctggccggcctgctctggtgcctggcctcgcgccgccgtgtatcgccccgccctgggcggcaaggctggcccggtcggcaccagttgcgtga
gcggaaagatggccgcttcccggccctgctgcagggagctcaaaatggaggacgcggcgctcgggagagcgggcgggtgagtcacccacac
aaaggaaaagggcctttccgtcctcagccgtcgcttcatgtgactccacggagtaccgggcgccgtccaggcacctcgattagttctcgagcttttg
gagtacgtcgtctttaggttgggggagggtttttatgcgatggagtttccccacactgagtgggtggagactgaagttaggccagcttggcacttg
atgtaattctccttggaatttgccctttttgagtttggatcttggttcattctcaagcctcagacagtggttcaaagttttttttcttccatttcaggtgtcgtgag
gatccgccaccatggtgagcaagggcgaggaggataacatggccatcatcaaggagttcatgcgcttcaaggtgcacatggagggctccgtgaa
cggccacgagttcgagatcgagggcgagggcgagggccgcccctacgagggcacccagaccgccaagctgaaggtgaccaagggcggccc
cctgcccttcgcctgggacatcctgtcccctcagttcatgtacggctccaaggcctacgtgaagcaccccgccgacatccccgactacctgaagct
gtccttccccgagggcttcaattgggagcgcgtgatgaacttcgaggacggcggcgtggtgaccgtgacccaggactcctccctgcaggacggc
gagttcatctacaaggtgaagctgcgcggcaccaacttcccctccgacggccccgtgatgcagtgtaggaccatgggctgggaggcctccactga
gcggatgtaccccgaggacggcgccctgaagggcgagatcaagcagaggctgaagctgaaggacggcggccactacgacgctgaggtgaag
accacctacaaggccaagaagcccgtgcagctgcccggcgcctacaacgtggacatcaagctggacatcctgtcccacaacgaggactacacc
atcgtggagcagtacgagcgcgccgagggccgccactccaccggcggcatggacgagctgtacaagggcagcggagaaggtagaggttctct
cctcacttgtggtgatgttgaagaaaaccctggtccaatgaccgagtacaagcccacggtgcgcctcgccacccgcgacgacgtccccagggcc

gtacgcaccctcgccgccgcgttcgccgactaccccgccacgcgccacaccgtcgatccggaccgccacatcgagcgggtcaccgagctgcaa
gaactcttcctcacgcgcgtcgggctcgacatcggcaaggtgtgggtcgcggacgacggcgccgccgtggcggtctggaccacgccggagag
cgtcgaagcgggggcggtgttcgccgagatcggcccgcgcatggccgagttgagcggttcccggctggccgcgcagcaacagatggaaggcc
tcctggcgccgcaccggcccaaggagcccgcgtggttcctggccaccgtcggagtctcgcccgaccaccagggcaagggtctgggcagcgcc
gtcgtgctccccggagtggaggcggccgagcgcgccgggggtgcccgccttcctggagacctccgcgccccgcaacctccccttctacgagcgg
ctcggcttcaccgtcaccgccgacgtcgaggtgcccgaaggaccgcgcacctggtgcatgacccgcaagcccggtgcctgaacgcgttaagtcg
acaatcaacctctggattacaaaatttgtgaaagattgactggtattcttaactatgttgctccttttacgctatgtggatacgctgctttaatgcctttgtatc
atgctattgcttcccgtatggctttcattttctcctccttgtataaatcctggttgctgtctctttatgaggagttgtggcccgttgtcaggcaacgtggcgtg
gtgtgcactgtgtttgctgacgcaacccccactggttggggcattgccaccacctgtcagctcctttccgggactttcgctttccccctccctattgcca
cggcggaactcatcgccgcctgccttgcccgctgctggacaggggctcggctgttgggcactgacaattccgtggtgttgtcggggaaatcatcgt
cctttccttggctgctcgcctgtgttgccacctggattctgcgcgggacgtccttctgctacgtcccttcggccctcaatccagcggaccttccttcccg
cggcctgctgccggctctgcggcctcttccgcgtcttcgccttcgccctcagacgagtcggatctccctttgggccgcctccccgcgtcgacttttaag
accaatgacttacaaggcagctgtagatcttagccacttttaaaagaaaaggggggactggaagggctaattcactcccaacgaagacaagatttc
accattatcgtttcagacccacctcccaaccccgaggggacccagagagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgt
tagagagataattagaattaatttgactgtaaacacaaagatattagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagttttaaa
attatgttttaaaatggactatcatatgcttaccgtaacttgaaagtatttcgatttcttggctttatatatcttgtggaaaggacgaaacaccggagacggt
tgtaaatgagcacacaaaatacacatgctaaaatattatattctatgacctttataaaatcaaccaaaatcttctttttaataactttagtatcaataattagaa
ttttatgttccttttgcaaacttttaataaaaatgagcaaaataaaaaaacgctagttttagtaactcgcgttgttttcttcacctttaataatagctactcca
ccacttgttcctaagcggtcagctcctgcttcaatcatttttgagcatcttcaaatgttctaactccaccagctgctttaactaaagcattgtctttaacaac
tgacttcattagtttaacatcttcaaatgttgcacctgattttgaaaatcctgttgatgttttaacaaattctaatccagcttcaacagctatttcacaagctttc
atgatttcttctttgttaataaacaattttccataatacatttaacaacatgtgatccagctgctttttttacagctttcatgtcttctaaaactaattcataatttt
gtcttttaatgcaccaatatttaataccatatcaatttctgttgcaccatctttaattgcttcagaaacttcgaatgcttttgtagctgttgtgcatgcacctag
aggaaaacctacaacatttgttattcctacatttgtgccttttaataattctttacaatagcttgttcaatatgaattaacacaaactgttgcaaaatcaaattc
aattgcttcatcacataattgtttaatttcagctttcgtagcatcttgttttaataatgtgtgatctatatatttgtttagtttcatttttttctcctatatattcatttttaa
ttttaattctttaataatttcgtctactttaactttagcgttttgaacagattcaccaacacctataaaataaatttttagtttaggttcagttccacttgggcgaa
cagcaaatcatgacttatcttctaaataaaatttttagtaagtcttgtcctggcatattatacattccatcgacgtctctgtttaagagctaggccaacatgag
gatcacccatgtctgcagggcctagcaagtttaaataaggctagtccgttatcaacttggccaacatgaggatcacccatgtctgcagggccaagtg
gcaccgagtcggtgctttttaagcttggcgcaccaagcttggcgtaactagatcttgagacactgctttttgcttgtactgggtctctctggttagacca
gatctgagcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgcttcaagtagtgtgtgcccgtctgtt
gtgtgactctggtaactagagatccctcagacccttttagtcagtgtggaaaatctctagcagtacgtatagtagttcatgtcatcttattattcagtatttat
aacttgcaaagaaatgaatatcagagagtgagaggaacttgtttattgcagcttataatggttacaaataaagcaatagcatcacaaatttcacaaataa
agcatttttttcactgcattctagttgtggtttgtccaaactcatcaatgtatcttatcatgtctggctctagctatcccgcccctaactccgcccgttccgcc
cattctccgccccatggctgactaatttttttttatttatgcagaggccgaggccgcctcggcctctgagctattccagaagtagtgaggaggctttttgg
aggcctagggacgtacccaattcgccctatagtgagtcgtattacgcgcgctcactggccgtcgttttacaacgtcgtgactgggaaaaccctggcg
ttacccaacttaatcgccttgcagcacatccccctttcgccagctggcgtaatagcgaagaggcccgcaccgatcgcccttcccaacagttgcgcag
cctgaatggcgaatgggacgcgccctgtagcggcgcattaagcgcggcgggtgtggtggttacgcgcagcgtgaccgctacacttgccagcgc
cctagcgcccgctcctttcgctttcttcccttcctttctcgccacgttcgccggctttccccgtcaagctctaaatcgggggctccctttagggttccgatt
tagtgctttacggcacctcgaccccaaaaaacttgattagggtgatggttcacgtagtgggccatcgccctgatagacggtttttcgccctttgacgtt
ggagtccacgttctttaatagtggactcttgttccaaactggaacaacactcaaccctatctcggtctattcttttgatttataagggattttgccgatttcg
gcctattggttaaaaaatgagctgatttaacaaaaatttaacgcgaattttaacaaaatattaacgcttacaatttaggtggcacttttcggggaaatgtgc
gcggaaccctatttgtttatttttctaaatacattcaaatatgtatccgctcatgagacaataaccctgataaatgcttcaataatattgaaaaaggaaga
gt3’

# Appendix C Screen candidates, sgRNA sequences, target gene activation and screen hits

1st column: sgRNA ID

2nd column: sgRNA target sequence

3rd column: target gene

4th column: target transcript

5th column: off-targets

6th column: number of counts average read 1 and read 2 plasmid library

7th column: number of cells expressing the sgRNA in 10X Genomics pooled CRISPRa dataset

8th column: number of cells with activation (FDR 10%, chapter 4)

9th column: effective / not effective sgRNA (more than 10 cells activated), classification chapter 4

10th column: fraction of variance explained by MOFA factor 3

11th column: hit rank

12th column: screen hit (yes / no)

13th column: mean log fold change expression of target gene in 10X Genomics pooled screen CRISPRa dataset

14th column: number of DE genes (FDR 10%) in 10X Genomics pooled screen CRISPRa dataset

15th column: number of DE genes recovered from ZGA signature (from Appendix I) in 10X Genomics pooled screen CRISPRa dataset

16th column: hit above background in differential gene expression rank

17th column: sgRNA used in other experiments

Appendix C: Ranked hit pooled sgRNA library, CRISPRa scRNA-seq dataset with three replicates merged, screen results and information of sgRNAs used for pilot and validation experiments

| | sgRNA ID | sgRNA target sequence | Target gene | Target transcript | Off-targets | Number of counts average read 1 and read 2 plasmid library | Number of cells expressing the sgRNA in 10X Genomics pooled CRISPRa dataset | Number of cells with activation (FDR 10%, chapter 4) | Effective/not effective (hence then 10 cells activated) classification in chapter 4 | Fraction of variance explained by MOFA factor 3 | Hit rank | Screen hit (Yes/No) | Mean log fold-change expression of target gene in 10X Genomics pooled screen CRISPRa dataset | Number of DE genes (FDR 10%) in 10X Genomics pooled screen CRISPRa dataset | Number of genes recovered from ECA signature (from Appendix I) in 10X Genomics pooled screen CRISPRa dataset | Hit above background in differential gene expression rank | sgRNA used in other experiments |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | CAACTTTCGTTGTAGAGAG | Bcl2b | NM_001001182 | No | 40 | 40 | 0 | Not effective | 0.218999133 | 584 | No | 0.001563828 | 0 | 23 | No |
| 2 | | GCCGCCGCCGCCGCCGCCG | Bcl2b | NM_001001182 | No | 0 | 0 | 0 | Not effective | 0 | 435 | No | -0.02512443 | 0 | 0 | No |
| 3 | | AGGAACTGTTCTTCTAAAGA | Bcl2b | NM_001001182 | No | 0 | 0 | 0 | Not effective | 0.006339015 | 435 | No | 0 | 0 | 0 | No |
| 4 | | GAGAACTGAGTTGCTTAGAG | Bcl2b | NM_001000081 | No | 29 | 29 | 0 | Not effective | 0.034459496 | 426 | No | 0 | 0 | 2 | No |
| 5 | | AGTTAGCGCGGCGGCGCTAGA | Bcl2b | NM_001000081 | No | 1343 | 1343 | 0 | Not effective | 0.004189015 | 88 | No | 0.00274793 | 0 | 2 | No |
| 6 | | CAAAGCGCTACCACAGAGTC | Mdc1 | NM_001010833 | No | 7304.5 | 6237.5 | 0 | Not effective | 0.351612396 | 396 | No | 0.055368243 | 0 | 1 | No |
| 7 | | AGCTGATTATGTGTCTACTT | Mdc1 | NM_001010833 | No | 5503 | 4703 | 0 | Not effective | 0.351612396 | 396 | No | 0.055368243 | 0 | 1 | No |
| 8 | | GGGCGCCTTGAAATCAAACT | Trf2 | NM_001013426 | No | 5348.5 | 5131 | 0 | Not effective | 0.659603098 | 513 | No | 0.09779173 | 0 | 1 | No |
| 9 | | GGGCGCTGGGAGGGCGGCAG | Trf2 | NM_001013426 | No | 13899 | 11719 | 0 | Not effective | 0.659603098 | 327 | No | -0.039192176 | 0 | 1 | No |
| 10 | | GCGCGCGCTCGTCGTGCGCC | Cnnbp | NM_001031772 | No | 3197 | 3134 | 0 | Not effective | 0.778049848 | 280 | No | -0.035795591 | 0 | 0 | No |
| 11 | | ACGAGTGACGTGTTCTCTAG | Ino80b | NM_001033941 | No | 3408 | 3134 | 0 | Not effective | 0.778049848 | 280 | No | -0.034892345 | 0 | 0 | No |
| 12 | | AAGTGCTTGCGCAGGAAGGC | Ino80b | NM_001033941 | No | 1134 | 1134 | 0 | Not effective | 0.778049848 | 135 | No | -0.01409346 | 0 | 0 | No |
| 13 | | GATGGTTAGGTGATTGGCTACT | Ino80b | NM_001033941 | No | 2225.5 | 1134 | 0 | Not effective | 0.778049848 | 135 | No | 0 | 0 | 0 | No |
| 14 | | GTCTTTCATGGTGCAGTCTTC | Ing2 | NM_001033427 | No | 783 | 783 | 0 | Not effective | 0.127741612 | 406 | No | 0.097594256 | 0 | 0 | No |
| 15 | | AGCTGATTATGTGTCTACTT | Ing2 | NM_001033427 | No | 1737 | 1737 | 0 | Not effective | 0.127741612 | 406 | No | 0 | 0 | 0 | No |
| 16 | | GAGGTGATTGCTTTAAAGAG | Ing2 | NM_001033427 | No | 595 | 595 | 0 | Not effective | 0.127741612 | 406 | No | 0 | 0 | 0 | No |
| 17 | | GGGCGCTTGAAATCAAACT | Trf1 | NM_001013426 | No | 5348.5 | 5131 | 0 | Not effective | 0.659603098 | 513 | No | 0.09779173 | 0 | 1 | No |
| 18 | | GGGCGCTGGGAGGGCGGCAG | Trf1 | NM_001013426 | No | 13899 | 11719 | 0 | Not effective | 0.659603098 | 327 | No | -0.039192176 | 0 | 1 | No |
| 19 | | AGCGCGCCCGCCATGATTTC | Cnnbp | NM_001031772 | No | 3197 | 3134 | 0 | Not effective | 0.778049848 | 280 | No | -0.035795591 | 0 | 0 | No |
| 20 | | GAGACTCCGCGTGAGTCTAG | Ino80b | NM_001033941 | No | 3408 | 3134 | 0 | Not effective | 0.778049848 | 280 | No | -0.034892345 | 0 | 0 | No |
| 21 | | CAAGCTCGGATTATCTAGAG | Ino80b | NM_001033941 | No | 1134 | 1134 | 0 | Not effective | 0.778049848 | 135 | No | -0.01409346 | 0 | 0 | No |
| 22 | | GTCGTCGCGGAGCGCGCAGCAG | Ino80b | NM_001033941 | No | 2225.5 | 1134 | 0 | Not effective | 0.778049848 | 135 | No | 0 | 0 | 0 | No |
| 23 | | GCGCGCGCGAGCGGCGGCAGCAG | Ino80b | NM_001033941 | No | 1737 | 1737 | 0 | Not effective | 0.127741612 | 406 | No | 0.097594256 | 0 | 0 | No |
| 24 | | GGGTGGCGGATCCCGCGAGCCC | Ino80b | NM_001033941 | No | 595 | 595 | 0 | Not effective | 0.127741612 | 406 | No | 0 | 0 | 0 | No |
| 25 | | GTACGAAGCGCAGTCTTCTAG | Ino80b | NM_001033941 | No | 783 | 783 | 0 | Not effective | 0.127741612 | 406 | No | 0 | 0 | 0 | No |
| 26 | | GAGGTGATTGCTTTAAAGAG | Ino80b | NM_001033941 | No | 1737 | 1737 | 0 | Not effective | 0.127741612 | 406 | No | 0 | 0 | 0 | No |
| 27 | | GGGCGCTGGGAGGGCGGCAG | Ino80b | NM_001033941 | No | 595 | 595 | 0 | Not effective | 0.127741612 | 406 | No | 0 | 0 | 0 | No |
| 28 | | GTACGAAGCGCAGTCTTCTAG | Ino80b | NM_001033941 | No | 783 | 783 | 0 | Not effective | 0.127741612 | 406 | No | 0 | 0 | 0 | No |
| 29 | | GTACGACGCACACTGAATCT | Ino80b | NM_001033941 | No | 3343 | 3343 | 0 | Not effective | 0.127741612 | 406 | No | 0 | 0 | 0 | No |
| 30 | | GCTCTTCTCGTTTCTTCGTC | Chd1 | NM_001042541 | Multi-mapping | 6097.5 | 5131 | 0 | Not effective | 0.659603098 | 513 | No | 0.09779173 | 0 | 1 | No |
| 31 | | GGGCAGAAGCAAGCATCGTTT | Chd1 | NM_001079830 | No | 6097.5 | 5131 | 0 | Not effective | 0.659603098 | 513 | No | 0.09779173 | 0 | 1 | No |
| 32 | | TCACAGCCCTTAGAGAAGG | Trim33 | NM_001008819 | No | 3153.5 | 2205 | 1 | Not effective | 0.132527146 | 371 | No | -0.018134035 | 0 | 2 | No |
| 33 | | GAGAGAGCGAAGAGAGGGCC | Trim33 | NM_001008819 | No | 7820.5 | 7820.5 | 0 | Not effective | 0.132527146 | 188 | No | -0.018134035 | 0 | 2 | No |
| 34 | | GAAGACTCTGTCTCAGCCC | Trim33 | NM_001079830 | No | 7820.5 | 7820.5 | 0 | Not effective | 0.132527146 | 188 | No | -0.018134035 | 0 | 2 | No |
| 35 | | GCTGCGCGCGCGCCATTGCT | Fosl2 | NM_001080921 | No | 15329 | 15329 | 0 | Not effective | 0.206082739 | 389 | No | -0.030806996 | 0 | 3 | No |
| 36 | | CTGATGAGCGCGAATAGCTT | Fosl2 | NM_001080921 | No | 3230 | 3230 | 0 | Not effective | 0.206082739 | 389 | No | -0.030806996 | 0 | 3 | No |
| 37 | | AGCTGCGAAACTTCGTTGAT | Arid4a | NM_001081195 | No | 1513.5 | 1513.5 | 0 | Not effective | 1.482406599 | 161 | No | -0.00670707 | 0 | 0 | No |
| 38 | | GCCCTCTCGTTTCTTCGTTC | Arid4a | NM_001081195 | No | 3144.5 | 3144.5 | 0 | Not effective | 1.482406599 | 161 | No | -0.00670707 | 0 | 0 | No |
| 39 | | CCGGATACACAGCTATCTC | Pbrm1 | NM_001081251 | No | 12030 | 12030 | 0 | Not effective | 1.208458827 | 141 | No | 0.073211931 | 0 | 1 | No |
| 40 | | GCCTCTCGTTTCTTCGTTC | Pbrm1 | NM_001081251 | No | 12784.5 | 12784.5 | 0 | Not effective | 1.208458827 | 141 | No | 0.073211931 | 0 | 1 | No |
| 41 | | GGTTCAACGTTACGCGCTCC | Chd2 | NM_001081251 | No | 11846.5 | 6625 | 0 | Not effective | 0.869584198 | 257 | No | -0.082010916 | 0 | 1 | No |
| 42 | | ATGTAGAGATGTGTCAATCAA | Chd2 | NM_001081345 | No | 7983.5 | 83 | 0 | Not effective | 0 | 436 | No | -0.07977005 | 0 | 0 | No |
| 43 (or Dux-sgRNA 1) | | GCTCTCCTCTCTCTCTCTCT | Dux | NM_001081954 | Multi-mapping | 3245.5 | 17 | 2 | Not effective | 0.11458555 | 436 | No | -0.00473486 | 0 | 0 | No |
| 44 (or Dux-sgRNA 2) | | GAGCAATGAGCAGTGGTCCC | Dux | NM_001081954 | Multi-mapping | 12099 | 548 | 0 | Not effective | 1.223254683 | 137 | No | -0.00473486 | 0 | 0 | No |
| 45 | | GGCCAATGAGCTCTTACTTA | Ubqln1 | NM_001083319 | No | 6518 | 548 | 0 | Not effective | 0.869871843 | 150 | No | 0.040022049 | 0 | 4 | No |
| 46 | | AGCCGCTGCTGCGCCGCCCT | Ankrd1b | NM_001081555 | No | 2178.5 | 111 | 0 | Not effective | 0.869871843 | 316 | No | -0.022836162 | 0 | 0 | No |
| 47 | | GGGCGCTGCGCGCGCGCCGC | Lin9 | NM_001101102 | No | 14189 | 657 | 0 | Not effective | 1.134655311 | 160 | No | 0.001853264 | 0 | 3 | No |
| 48 | | GAACTGCAGCGCGGTTCCCG | Lin9 | NM_001101102 | No | 14189 | 657 | 0 | Not effective | 1.134655311 | 160 | No | -0.004189634 | 0 | 3 | No |
| 49 | | GTGGCGCGCGCGAGCGAGCC | Lin9 | NM_001101102 | No | 10460.5 | 841 | 0 | Not effective | 1.166890634 | 148 | No | -0.009952434 | 0 | 3 | No |
| 50 | | GGCTAGCGCCATTGTTTGTG | Fic1 | NM_001110329 | No | 3766 | 214 | 0 | Not effective | 0.894839442 | 202 | No | -0.011930782 | 0 | 0 | No |
| 51 | | TCCTCTCTCTCTCTCTCTCT | Fic1 | NM_001110329 | No | 7716 | 214 | 0 | Not effective | 0.894839442 | 202 | No | -0.011930782 | 0 | 0 | No |
| 52 | | TTGCAGCCAATGCATACTGC | Sin3a | NM_001100351 | No | 5466 | 155 | 0 | Not effective | 1.026820071 | 183 | No | -0.008234015 | 0 | 0 | No |
| 53 | | TTGCAGCCAGTCTTACTCG | Sin3a | NM_001100351 | No | 5466 | 2270 | 20 | Not effective | 1.125816294 | 134 | No | 0.073636819 | 0 | 1 | No |
| 54 | | CGCGCACTAGTCGCTCGCCT | Brd3 | NM_001113574 | No | 24783 | 1189 | 0 | Not effective | 0.566350546 | 331 | No | -0.026865887 | 0 | 4 | No |
| 55 | | GACATGTTACTCCGTCGCCT | Brd3 | NM_001113574 | No | 7142.5 | 476 | 3 | Not effective | 0.566350546 | 331 | No | -0.026865887 | 0 | 4 | No |
| 56 | | GCAGAGCTGTATCCGCCAG | Brd3 | NM_001114140 | No | 6975.5 | 148 | 0 | Not effective | 0.819473497 | 228 | No | -0.004893959 | 0 | 1 | No |
| 57 | | GAGCTCCGCGCGGAGGCAGAGC | Trrc18 | NM_001122730 | No | 6975.5 | 337 | 10 | Not effective | 0.819473497 | 69 | No | -0.004893959 | 14 | 1 | No |
| 58 | | AAGTGTGAGGCGCCTCGCGG | Trrc18 | NM_001122730 | No | 4471 | 337 | 3 | Not effective | 1.149896162 | 28 | No | 0.035038002 | 27 | 31 | No |
| 59 | | GGCTAGCGCCATTGTTTGTG | Sin3 | NM_001122786 | No | 2739.5 | 221 | 1 | Not effective | 0.272318488 | 379 | No | -0.011930782 | 0 | 0 | No |
| 60 | | GATTGGCTGTGTGCAGTAC | Sin3 | NM_001122786 | No | 7716 | 0 | 0 | Not effective | 0 | 379 | No | 0.010423109 | 0 | 0 | No |
| 61 | | TTGCAGCCAGTCTTACTCG | Sin3 | NM_001122786 | No | 5466 | 52 | 0 | Not effective | 1.125816294 | 134 | No | 0.073636819 | 0 | 1 | No |
| 62 | | CTCCTATTCGCGCGCCGCC | Sin3a | NM_001122802 | No | 13599 | 0 | 0 | Not effective | 1.049496808 | 63 | No | -0.037634112 | 0 | 0 | No |
| 63 | | TGCGCACGAAGCGCGCAAGG | Gm41 | NM_001122802 | No | 13099.5 | 783 | 0 | Not effective | 1.159096173 | 155 | No | -0.025961837 | 0 | 0 | No |
| 64 | | GCTGAAAGGGAGGGAGAAGG | Gm41 | NM_001122802 | No | 2098.5 | 66 | 0 | Not effective | 0 | 436 | No | -2.79E-04 | 0 | 0 | No |
| 65 | | TGAAATGAAAGATGGATTCT | Tgs3 | NM_001127260 | No | 7903 | 48 | 0 | Not effective | 0.001518654 | 409 | No | -2.79E-04 | 0 | 0 | No |
| 66 | | GCTGCTCTCTCTCTCTCTCT | Tgs3 | NM_001127260 | No | 6435.5 | 0 | 0 | Not effective | 0.001518654 | 409 | No | -2.79E-04 | 0 | 0 | No |
| 67 | | GTTCTTCTTCTGCTGCCGCC | Cec1 | NM_001131151 | No | 8353 | 545 | 0 | Not effective | 1.140933274 | 157 | No | -0.007137006 | 163 | 1 | No |
| 68 | | GCTCGCGCGAAGCTCTCTG | Mav | NM_001146176 | No | 6435.5 | 545 | 0 | Not effective | 0.419574091 | 157 | No | -0.007137006 | 163 | 1 | No |
| 69 | | ACTAGAGTTCTCAGCTACCC | Mav | NM_001146176 | No | 6435.5 | 545 | 0 | Not effective | 0.419574091 | 157 | No | -0.007137006 | 163 | 1 | No |
| 70 | | AGTCGCGCGAAGCTCTCTG | Mav | NM_001146176 | No | 8353 | 545 | 0 | Not effective | 0.419574091 | 157 | No | -0.007137006 | 163 | 1 | No |
| 71 | | GGTGCGCCGCCTGCCTCCGGA | Ivl | NM_001159386 | No | 12306.5 | 588 | 0 | Not effective | 0.800072018 | 334 | No | -0.019157559 | 0 | 0 | No |
| 72 | | GCGCCGCCGCCTGCCTCCGGA | Ivl | NM_001159386 | No | 11956.5 | 588 | 0 | Not effective | 0.800072018 | 334 | No | -0.019157559 | 0 | 0 | No |
| 73 | | GAGCCGCAGCCCCCTCGCTG | Srl1 | NM_001159588 | No | 11956.5 | 409 | 25 | Not effective | 0.121905703 | 41 | No | 0.123003206 | 148 | 88 | No |
| 74 | | TGACTTCCGCGCGAGGAGCGCC | Pes252 | NM_001163556 | No | 6333 | 409 | 73 | Effective | 1.733335325 | 41 | Yes | 0.123003206 | 148 | 88 | No |
| 75 | | CGCGCGAGGAAACGCAGCGC | Pes252 | NM_001163556 | No | 7863 | 216 | 47 | Effective | 0.312049819 | 375 | Yes | 0.202229375 | 15 | 59 | No |
| 76 | | GCTCTCCTCCTCAAACTCG | Srl1 | NM_001163632 | No | 7504 | 216 | 13 | Not effective | 0.312049819 | 375 | No | 0.202229375 | 15 | 59 | No |
| 77 | | GCTCTCCTCCTCAAACTCG | Srl1 | NM_001163632 | No | 7504 | 615 | 0 | Not effective | 1.140841118 | 75 | No | -0.007074898 | 7 | 3 | No |
| 78 | | AGCCGCGCCTGCGCGACGCGCC | Pes252 | NM_001164597 | No | 13076 | 615 | 0 | Not effective | 0.91594204 | 236 | No | -0.035833696 | 10 | 5 | No |
| 79 | | TTTCTGCGAGCGCGCGCGCGC | Pes252 | NM_001164597 | No | 13076 | 615 | 0 | Not effective | 0.91594204 | 236 | No | -0.035833696 | 10 | 5 | No |
| 80 | | AAGACGACGCGCCGCCGCTGAC | Hmgn1 | NM_001166545 | No | 11687.5 | 753 | 0 | Not effective | 1.314827804 | 385 | No | 0.009237327 | 5 | 0 | No |
| 81 | | GTCGCCGCCGCCGCCGCCGCC | Hmgn1 | NM_001166545 | No | 11687.5 | 1505 | 0 | Not effective | 1.314827804 | 109 | No | -0.010323947 | 0 | 0 | No |
| 82 | | CTAGTCTTAGCGGAGTCGCA | Phf14 | NM_001168382 | No | 6975.5 | 64 | 0 | Not effective | 0.096277448 | 415 | No | -0.014795419 | 1 | 1 | No |
| 83 | | CTAAGATAAGTATTTACTTC | Phf14 | NM_001168382 | No | 676 | 64 | 0 | Not effective | 0.096277448 | 415 | No | -0.014795419 | 1 | 1 | No |
| 84 | | TACTATGGAAATGAGAGTAC | Phf14 | NM_001168382 | No | 6975.5 | 70 | 0 | Not effective | 0 | 436 | No | -0.023754743 | 0 | 0 | No |
| 85 | | CTAAGATAAGTATTTACTTC | Phf14 | NM_001168382 | No | 676 | 27 | 0 | Not effective | 0.207554474 | 387 | No | 0.311118492 | 0 | 0 | No |
| 86 | | TTTCGTTAATTCTTCTTACT | Zfp37 | NM_001186501 | No | 1669.5 | 125 | 16 | Effective | 0 | 436 | No | 0.218433188 | 6 | 6 | No |
| 87 | | GGCTGCTGCCGCAGCGCTGC | Zfp37 | NM_001186501 | No | 1669.5 | 1497 | 1 | Not effective | 0.608563939 | 252 | No | 0.061339864 | 2 | 2 | No |
| 88 | | GCGCTGCCGCAGCGCTGC | Mta3 | NM_001172854 | No | 22774.5 | 1497 | 1 | Not effective | 2.380563939 | 11 | No | 0.061339864 | 2 | 2 | No |
| 89 | | AAGGAAGAGCGTGAGGAGGC | Ythdf1 | NM_001171147 | No | 13945 | 42 | 4 | Not effective | 0 | 436 | No | 0.020373942 | 6 | 6 | No |
| 90 | | CGGCTGCGTGGGACCCCAC | Ythdf1 | NM_001171147 | No | 21642 | 608 | 2 | Not effective | 2.227339854 | 12 | No | 0.087632867 | 32 | 30 | No |
| 91 | | AGTCGTTGAGGGGCGGCGAGG | Smarca4 | NM_001174078 | No | 4888 | 229 | 53 | Not effective | 0.827986171 | 377 | No | -0.008772135 | 5 | 1 | No |
| 92 | | GGCGCTGCGCGCGCGCGCGC | Smarca4 | NM_001174078 | No | 4888 | 55 | 2 | Not effective | 1.19421139 | 59 | No | -0.008772135 | 5 | 1 | No |
| 93 | | AAGACGCCGATAGCCTGCAT | Dnmt1 | NM_001190432 | No | 12152.5 | 755 | 0 | Not effective | 1.260482356 | 122 | No | -0.044728882 | 27 | 41 | No |
| 94 | | AGCATGCCCGCCTCTTCTCC | Dnmt1 | NM_001190432 | No | 12152.5 | 755 | 0 | Not effective | 1.260482356 | 122 | No | -0.044728882 | 1 | 1 | No |
| 95 | | ATTGTAGTTCTTCGCTTCTCG | Jarid2 | NM_001205044 | No | 4381 | 257 | 0 | Not effective | 0.758413555 | 285 | No | -0.039090423 | 0 | 18 | No |

**Appendix C: Related to pooled sgRNA library, CRISPRa scRNA-seq dataset with three replicates merged, screen results and information of sgRNAs used for pilot and validation experiments**

| sgRNA ID | sgRNA target sequence | Target gene | Target transcript | Off-targets | Number of counts average read 1 and read 2 plasmid library | Number of cells expressing the sgRNA in 10X Genomics pooled CRISPRa dataset | Number of cells with activation (FDR 10%, chapter 4) | Effective/not effective (more than 10 cells activated) classification in chapter 4 | Fraction of variance explained by MOFA factor 3 | Hit rank | Screen hit (Yes/No) | Mean log fold change expression of target gene in 10X Genomics pooled screen CRISPRa dataset | Number of DE genes (FDR 10%) in 10X Genomics pooled screen CRISPRa dataset | Number of genes recovered from 2GA signature (from Appendix I) in 10X Genomics pooled screen CRISPRa dataset | Hit above background in differential gene expression rank | sgRNA used in other experiments |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 96 | GTTGAGTGACAGGAGGCGTG | Jarid2 | NM_001205044 | No | 4128 | 253 | 0 | Not effective | 1.455840402 | 70 | No | -0.122443927 | 1 | 29 | It is not a hit | No |
| 97 | CAATTGAGAGGAGCAGAGCT | Eya1 | NM_001252192 | No | 2330 | 131 | 8 | Not effective | 0.76100889 | 284 | No | 0.069277357 | 10 | 25 | It is not a hit | No |
| 98 | CTGTTTATTCAGATATCATG | Eya1 | NM_001252192 | No | 1068.5 | 85 | 1 | Not effective | 0.119425413 | 399 | No | 0.026497809 | 1 | 27 | It is not a hit | No |
| 99 | AGCTTCAAGAAAGGACATGG | Shmt2 | NM_001252316 | No | 1639 | 187 | 0 | Not effective | 0.454783109 | 355 | No | -0.018335403 | 3 | 16 | It is not a hit | No |
| 100 | TCCCTGCGCATGCGCAGTAC | Shmt2 | NM_001252316 | No | 5179 | 449 | 0 | Not effective | 0.900775514 | 240 | No | -0.051965785 | 1 | 18 | It is not a hit | No |
| 101 | GAAAGCCTGGAGGTAGGCAG | Mlt10 | NM_001252561 | No | 2761 | 97 | 7 | Not effective | 0 | 436 | No | 0.209817774 | 31 | 17 | It is not a hit | No |
| 102 | GCGGAGTCGGCTGCTCAGCG | Mlt10 | NM_001252561 | No | 14325 | 798 | 16 | Effective | 1.257676686 | 125 | No | 0.079467068 | 6 | 37 | It is not a hit | No |
| 103 | GTGCTACGCGGGTCCGGCGG | Patz1 | NM_001253691 | No | 10380.5 | 337 | 104 | Effective | 2.769273513 | 4 | Yes | 0.34459421 | 160 | 53 | Yes | Validated by individual CRISPRa + bulk RNA-seq |
| 104 | GTTAAGGGCGGCAATGTGGG | Patz1 | NM_001253691 | No | 9366 | 167 | 2 | Not effective | 1.550852001 | 57 | No | -0.020654858 | 1 | 24 | It is not a hit | No |
| 105 | TGACCAAGCAGCCCGGTCCC | Tdrd3 | NM_001253755 | No | 15435.5 | 932 | 0 | Not effective | 0.603992816 | 325 | No | -0.005654853 | 10 | 18 | It is not a hit | No |
| 106 | TTCTAAGTTGCACACACATG | Tdrd3 | NM_001253755 | No | 1804.5 | 166 | 0 | Not effective | 1.438816742 | 74 | No | 0.008174321 | 2 | 45 | It is not a hit | No |
| 107 | AAGCAGAAAGAAGTTAATGC | Hdac9 | NM_001271386 | No | 748 | 29 | 0 | Not effective | 0 | 436 | No | -0.001150296 | 0 | 12 | It is not a hit | No |
| 108 | GTTCTGCTGAGACAATAAGA | Hdac9 | NM_001271386 | No | 1846 | 99 | 0 | Not effective | 0.109191695 | 404 | No | -0.001150296 | 0 | 16 | It is not a hit | No |
| 109 | CCCGCCGGACGCGAGGACCG | Trim24 | NM_001272064 | No | 15807 | 600 | 0 | Not effective | 1.492421423 | 64 | No | -0.098884924 | 22 | 27 | It is not a hit | No |
| 110 | CTGCAGAAGCCCGCGAGGCT | Trim24 | NM_001272064 | No | 11996.5 | 478 | 0 | Not effective | 1.259810689 | 123 | No | 0.011914068 | 16 | 33 | It is not a hit | No |
| 111 | CTTCCGGTCCCGCCGGGCGC | Chd7 | NM_001277149 | No | 32156.5 | 1314 | 0 | Not effective | 0.893553689 | 244 | No | 0.094560703 | 21 | 35 | It is not a hit | No |
| 112 | GGGCGGCGCGGGAGAGAGC | Chd7 | NM_001277149 | No | 11081 | 339 | 6 | Not effective | 1.730283821 | 42 | Yes | 0.069248714 | 9 | 40 | Yes | No |
| 113 | AGCATTGCAGCAGACACAAG | Terf2 | NM_001286200 | No | 2381.5 | 130 | 0 | Not effective | 0.06382693 | 417 | No | -0.002610996 | 17 | 18 | It is not a hit | No |
| 114 | TGAAGACTACGCTGCGGACT | Terf2 | NM_001286200 | No | 7245 | 175 | 0 | Not effective | 0.007056466 | 434 | No | -0.030866715 | 0 | 16 | It is not a hit | No |
| 115 | TCTCTGTCTCTTTGCTACTG | Mier1 | NM_001286221 | Multi-mapping | 3842.5 | 233 | 0 | Not effective | 1.029259287 | 190 | No | -0.070543308 | 0 | 25 | It is not a hit | No |
| 116 | TGTGAGACTGAGTGAGCTTG | Mier1 | NM_001286221 | No | 2412.5 | 167 | 0 | Not effective | 0.702114674 | 301 | No | -0.058035127 | 0 | 14 | It is not a hit | No |
| 117 | AAAGCAATTACGCAAGTTGG | Nfx1 | NM_001290449 | No | 1674.5 | 201 | 0 | Not effective | 0.197838212 | 391 | No | 0.004240335 | 1 | 23 | It is not a hit | No |
| 118 | CGCCATTTCACTGTGTGCTC | Nfx1 | NM_001290449 | No | 7527 | 294 | 0 | Not effective | 0.830011683 | 264 | No | -0.048965099 | 3 | 24 | It is not a hit | No |
| 119 | CGTGAAAGTGACTGCGGAGC | Cux1 | NM_001291240 | No | 5546.5 | 242 | 0 | Not effective | 1.340405161 | 101 | No | -0.017214876 | 0 | 17 | It is not a hit | No |
| 120 | GAACCGCCACACCGGGTAGC | Cux1 | NM_001291240 | No | 14587 | 950 | 1 | Not effective | 0.940602838 | 227 | No | -0.027987075 | 3 | 14 | It is not a hit | No |
| 121 | CGAGAGGCGCGGGAGATGAG | Tfdp1 | NM_001291765 | No | 4776.5 | 226 | 0 | Not effective | 1.089568816 | 170 | No | 0.03109993 | 0 | 21 | It is not a hit | No |
| 122 | GGTGCTACCGCCCTGGCGAG | Tfdp1 | NM_001291765 | No | 11226.5 | 582 | 0 | Not effective | 0.972510154 | 213 | No | -0.006681824 | 633 | 30 | It is not a hit | No |
| 123 | CAACCTCCTGCCTGGCCAT | Hmgb3 | NM_001293623 | No | 14922.5 | 996 | 28 | Effective | 1.417183181 | 80 | No | -0.001289531 | 2 | 24 | It is not a hit | No |
| 124 | CATTCAAACAGAACAACCCG | Hmgb3 | NM_001293623 | No | 2958 | 173 | 55 | Effective | 1.113112841 | 165 | No | 0.452350647 | 5 | 32 | It is not a hit | No |
| 125 | GGCAGCTTCATCGTCATTCG | Preb | NM_001294302 | No | 6880 | 373 | 92 | Effective | 0.845943871 | 260 | No | 0.292933671 | 1 | 15 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 126 | TAGGACCGGATCCGACGCC | Preb | NM_001294302 | No | 8894 | 355 | 66 | Effective | 2.006948667 | 20 | Yes | 0.223845425 | 2 | 33 | Yes | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 127 | CCCGACTAGTCAGTGGGCGG | Usp3 | NM_001302116 | No | 2587 | 246 | 0 | Not effective | 1.440335271 | 72 | No | -0.034388642 | 2 | 37 | It is not a hit | No |
| 128 | CGTAACCAGCAGGGTGGGCG | Usp3 | NM_001302116 | No | 7606.5 | 405 | 0 | Not effective | 1.78065436 | 37 | Yes | -0.020422553 | 5 | 25 | It is not a hit | No |
| 129 | CAGCTAACGGATGTGGCATG | Phf1 | NM_001302397 | No | 3038.5 | 174 | 1 | Not effective | 1.063164394 | 179 | No | 0.009136807 | 3 | 18 | It is not a hit | No |
| 130 | GGAGCTTGGAGTGGCCACGG | Phf1 | NM_001302397 | No | 7215.5 | 370 | 1 | Not effective | 1.856609485 | 25 | Yes | 0.015536536 | 0 | 28 | It is not a hit | No |
| 131 | GCGACACAAGGGACCATGG | Ing1 | NM_001302458 | No | 4069 | 301 | 0 | Not effective | 0.552937324 | 338 | No | -0.048867231 | 6 | 28 | It is not a hit | No |
| 132 | TCCCGGTGCTGGGTTTAGAA | Ing1 | NM_001302458 | No | 2064 | 153 | 8 | Not effective | 1.624884752 | 49 | No | 0.100029856 | 1 | 20 | It is not a hit | No |
| 133 | CGCAGCCAATCAGCAACGGC | Cic | NM_001302811 | No | 11886.5 | 658 | 0 | Not effective | 1.112595342 | 166 | No | -0.050557928 | 2 | 28 | It is not a hit | No |
| 134 | TAAACCAGAGTAGAGAAGTT | Cic | NM_001302811 | No | 844.5 | 42 | 0 | Not effective | 0.19030871 | 393 | No | -0.108885177 | 7 | 13 | It is not a hit | No |
| 135 | GCCGCCCGGGAGGAGCGGGAG | Ubtf | NM_001302954 | No | 10494.5 | 398 | 0 | Not effective | 0.987169567 | 204 | No | -0.090666982 | 5 | 24 | It is not a hit | No |
| 136 | GCCTTGGGCCGCCGCCTCC | Ubtf | NM_001302954 | No | 30076 | 1298 | 0 | Not effective | 1.130628681 | 162 | No | -0.036790786 | 6 | 25 | It is not a hit | No |
| 137 | CGCGGGAAACTTGAACCTTA | Conk2b | NM_001303446 | No | 3123.5 | 191 | 0 | Not effective | 0.72232616 | 294 | No | -0.151182139 | 0 | 27 | It is not a hit | No |
| 138 | GCACGCGATGCTTTACTACG | Conk2b | NM_001303446 | No | 4529.5 | 250 | 1 | Not effective | 0.873683993 | 252 | No | -0.011449413 | 1 | 15 | It is not a hit | No |
| 139 | CTCCACTTTCAAGGATTGTA | Cbx1 | NM_007622 | No | 1571.5 | 88 | 7 | Not effective | 2.981382043 | 1 | Yes | 0.124524294 | 6 | 39 | Yes | No |
| 140 | TGGGCGGCGGGGATTCGCGG | Cbx1 | NM_007622 | No | 15233 | 701 | 4 | Not effective | 0.731327787 | 290 | No | 0.020158044 | 2 | 19 | It is not a hit | No |
| 141 | GCTGCGTGCTCCGGGAATCC | Cbx2 | NM_007623 | No | 12126 | 619 | 62 | Effective | 1.373407186 | 93 | No | 0.179323118 | 11 | 23 | It is not a hit | No |
| 142 | GGGGGCTCTCTGCACTGTGT | Cbx2 | NM_007623 | No | 9838 | 618 | 90 | Effective | 0.730738111 | 291 | No | 0.222913908 | 37 | 19 | It is not a hit | No |
| 143 | CCCGCCGGGGAACGCGTCCTG | Cbx3 | NM_007624 | No | 22436 | 658 | 0 | Not effective | 0.70956507 | 297 | No | -0.040070993 | 15 | 20 | It is not a hit | No |
| 144 | GCGCGCTCCGCAGCGCAGAG | Cbx3 | NM_007624 | No | 9735 | 137 | 6 | Not effective | 0.825400098 | 429 | No | 0.116516585 | 16 | 27 | It is not a hit | No |
| 145 | CGCGCGAGATCTGCTGGCCT | Dhx9 | NM_007842 | No | 10383 | 735 | 0 | Not effective | 0.862847209 | 258 | No | -0.073792962 | 4 | 27 | It is not a hit | No |
| 146 | GTCGGCGACTCGCGCCTGCG | Dhx9 | NM_007842 | No | 23187.5 | 1405 | 0 | Not effective | 1.064807582 | 178 | No | -0.086306822 | 1 | 20 | It is not a hit | No |
| 147 | AGAGAACAGCTTTCCGTGGC | Dnajc1 | NM_007869 | No | 4882 | 274 | 103 | Effective | 0.819516118 | 268 | No | 0.409551378 | 1 | 19 | It is not a hit | No |
| 148 | AGCTGCACGTGTGTCAGCGC | Dnajc1 | NM_007869 | No | 8481 | 419 | 54 | Effective | 0.928275175 | 232 | No | 0.210829631 | 1 | 18 | It is not a hit | No |
| 149 | CGTTACTGAGGACAGCCGCG | E2f5 | NM_007892 | No | 12390.5 | 702 | 0 | Not effective | 0.87436453 | 211 | No | -0.030613926 | 2 | 18 | It is not a hit | No |
| 150 | GGCAGTCGGGTTCCCTCAGG | E2f5 | NM_007892 | No | 11032.5 | 492 | 305 | Effective | 0.491402299 | 350 | No | 0.764340012 | 64 | 30 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 151 | AGGCGCTTGATAGTGCTGGG | Ezh2 | NM_007971 | No | 3853.5 | 280 | 0 | Not effective | 1.060772014 | 180 | No | -0.074843865 | 8 | 23 | It is not a hit | No |
| 152 | TTCGGAGCGAGCTCCAGCCC | Ezh2 | NM_007971 | No | 9162 | 439 | 206 | Effective | 0.922400052 | 233 | No | 0.461490981 | 59 | 35 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 153 | CTGCAACCAAACCGCTCC | Gata3 | NM_008091 | No | 17568 | 92 | 19 | Effective | 2.507599639 | 7 | Yes | 0.28737548 | 224 | 38 | Yes | No |
| 154 | TGGCGCGACGCAACTTAAGG | Gata3 | NM_008091 | No | 3681 | 25 | 2 | Not effective | 0.018285688 | 431 | No | 0.070142849 | 8 | 21 | It is not a hit | No |
| 155 | CTTTGTGGATGGGAGCAGGG | Nr3c1 | NM_008173 | Multi-mapping | 3103.5 | 242 | 0 | Not effective | 1.33050009 | 103 | No | -0.037875763 | 27 | 28 | It is not a hit | No |
| 156 | GCCTATTAAATAAACCCTTT | Nr3c1 | NM_008173 | No | 960.5 | 47 | 0 | Not effective | 0 | 436 | No | 0.021818189 | 0 | 16 | It is not a hit | No |
| 157 | ACGGGTAGTCACACAGAGTC | Hdac2 | NM_008229 | No | 6022.5 | 299 | 0 | Not effective | 1.653233033 | 47 | No | -0.102417399 | 0 | 32 | It is not a hit | No |
| 158 | AGTCCGCTTGTGCGCACCTC | Hdac2 | NM_008229 | No | 9239 | 636 | 1 | Not effective | 0.28755509 | 378 | No | -0.064021812 | 1 | 15 | It is not a hit | No |
| 159 | ATGTACTGCTCGGCCCTGAT | Hmgb2 | NM_008252 | No | 3644 | 304 | 41 | Effective | 1.060127565 | 181 | No | 0.068661055 | 1 | 19 | It is not a hit | No |
| 160 | CTCTCTAGACCGGTTCCAAC | Hmgb2 | NM_008252 | No | 6766 | 481 | 310 | Effective | 1.246145471 | 126 | No | 0.258805366 | 13 | 33 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 161 | CTCTGATTGGTGAGCAGCCC | Hnf1 | NM_008296 | No | 6722 | 388 | 5 | Not effective | 1.59831724 | 51 | No | 0.039898658 | 35 | 52 | It is not a hit | No |
| 162 | TGGCCGGCCAAGTAGGCGAG | Hnf1 | NM_008296 | No | 2744 | 73 | 1 | Not effective | 0.042448583 | 425 | No | -0.065830019 | 2 | 12 | It is not a hit | No |
| 163 | CGCGGTGCGCGTGTGAGCGA | Smad1 | NM_008539 | No | 5897.5 | 181 | 52 | Effective | 0.915788728 | 237 | No | 0.356630794 | 81 | 35 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 164 | GAAGGCGGGGCCGGTAATT | Smad1 | NM_008539 | No | 8302 | 357 | 58 | Effective | 1.292164844 | 115 | No | 0.209874455 | 39 | 41 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 165 | GGACACTGAGCATGCGGGCG | Smad4 | NM_008540 | No | 12243 | 843 | 0 | Not effective | 0.976508254 | 210 | No | -0.008980758 | 1 | 21 | It is not a hit | No |
| 166 | TGGCAGCAACAACACGGCCC | Smad4 | NM_008540 | No | 8705 | 476 | 1 | Not effective | 0.823922659 | 267 | No | 0.033959779 | 1 | 24 | It is not a hit | No |
| 167 | AGCCCGCCGCCGCAGAACCCT | Mybl2 | NM_008852 | No | 12488.5 | 755 | 3 | Not effective | 1.353219298 | 100 | No | 0.147455118 | 76 | 28 | It is not a hit | No |
| 168 | CTATCTCCGCCAAGTGCGT | Mybl2 | NM_008852 | No | 10066.5 | 493 | 2 | Not effective | 1.16142519 | 151 | No | 0.025807443 | 1 | 19 | It is not a hit | No |
| 169 | CGCCCAGCACTACGGGTCCC | Ncoa3 | NM_008879 | Multi-mapping | 16155.5 | 765 | 0 | Not effective | 1.050884251 | 185 | No | -0.029706071 | 33 | 31 | It is not a hit | No |
| 170 | GGAGGTGAAGAGGAGGTTCT | Ncoa3 | NM_008879 | No | 2795.5 | 78 | 29 | Effective | 0 | 436 | No | 0.595703488 | 37 | 24 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 171 | GAGCAGAGCCCGCGGCAGCG | Rb1 | NM_009029 | No | 12479.5 | 808 | 29 | Effective | 0.649408942 | 310 | No | 0.127490622 | 16 | 25 | It is not a hit | No |
| 172 | TCGCGGGCGCGCGGGCGTCAG | Rb1 | NM_009029 | No | 14988.5 | 689 | 5 | Not effective | 1.396330015 | 85 | No | 0.069370297 | 6 | 45 | It is not a hit | No |
| 173 | ATTGGTCGGTGTTGGGCGA | Rbbp4 | NM_009030 | No | 5090 | 391 | 0 | Not effective | 0.712724688 | 296 | No | -0.065730628 | 0 | 12 | It is not a hit | No |
| 174 | TGGGCGGAGCATCGGCCTCG | Rbbp4 | NM_009030 | No | 13499 | 949 | 0 | Not effective | 1.379726098 | 90 | No | -0.097038717 | 0 | 24 | It is not a hit | No |
| 175 | GATGTCACCGCGGCGGGAC | Rela | NM_009045 | No | 11507 | 606 | 14 | Effective | 1.761580998 | 39 | Yes | 0.094854378 | 11 | 37 | Yes | No |
| 176 | TGCGGCCATTCGCCCTGCGC | Rela | NM_009045 | No | 30258 | 1293 | 2 | Not effective | 0.888621176 | 245 | No | 0.011363425 | 8 | 31 | It is not a hit | No |
| 177 | GGAGCTCGGTGAAGTCGAGA | Rfx1 | NM_009055 | No | 2621 | 148 | 0 | Not effective | 0.845273222 | 261 | No | 3.54E-05 | 0 | 25 | It is not a hit | No |
| 178 | GGGGGCTTCCAGCATGCTAT | Rfx1 | NM_009055 | No | 6000.5 | 306 | 0 | Not effective | 0.682744132 | 303 | No | -0.026287621 | 1 | 21 | It is not a hit | No |
| 179 | CCGCCCAGCCAACCCTCTTC | Smarcc1 | NM_009211 | No | 21318 | 1201 | 0 | Not effective | 1.240944088 | 128 | No | -0.033652719 | 6 | 44 | It is not a hit | No |
| 180 | CGGCGAGCCTCAGTCCGTAG | Smarcc1 | NM_009211 | No | 10033.5 | 474 | 84 | Effective | 1.016513254 | 195 | No | 0.239582788 | 9 | 39 | It is not a hit | No |
| 181 | CGCCTGGGCTCGGATCGACA | Terf1 | NM_009352 | No | 10116.5 | 605 | 30 | Effective | 1.449995579 | 71 | No | 0.050622108 | 1 | 28 | It is not a hit | No |
| 182 | TTTAGCAGTTCTCAGCCAAT | Terf1 | NM_009352 | No | 1989.5 | 186 | 0 | Not effective | 1.19108105 | 143 | No | -0.106884343 | 6 | 25 | It is not a hit | No |
| 183 | CTGTGGGAAGCCCGGACTTC | Tfam | NM_009360 | No | 9217 | 487 | 0 | Not effective | 1.367956068 | 95 | No | -0.031624218 | 1 | 32 | It is not a hit | No |
| 184 | GGCGGAGCCTGAGGCTAGGG | Tfam | NM_009360 | No | 5754.5 | 226 | 6 | Not effective | 0.892306591 | 407 | No | 0.041451308 | 25 | 22 | It is not a hit | No |
| 185 | TGGCTGGCGTTTCCCAGGCC | Usf1 | NM_009480 | No | 9901 | 663 | 4 | Not effective | 0.769855999 | 282 | No | -0.038130842 | 13 | 21 | It is not a hit | No |
| 186 | TGGGCACGGAGCCTTTGTTT | Usf1 | NM_009480 | No | 3656 | 200 | 0 | Not effective | 0.515634299 | 346 | No | -0.039843978 | 15 | 24 | It is not a hit | No |
| 187 | CGCCGGGTTCCCTTAGCAAC | Xrcc1 | NM_009532 | No | 9586 | 527 | 3 | Not effective | 1.597265217 | 52 | No | 0.117426054 | 3 | 23 | It is not a hit | No |
| 188 | GTTGGGCCTCTCGCGCAGCG | Xrcc1 | NM_009532 | No | 15916 | 906 | 0 | Not effective | 1.244169357 | 127 | No | 0.100100072 | 5 | 37 | It is not a hit | No |
| 189 | ACGTAGGAACTGATGGGTCT | Dnajc2 | NM_009584 | No | 3157 | 189 | 2 | Not effective | 0.233034032 | 380 | No | 0.05215876 | 2 | 13 | It is not a hit | No |
| 190 | GCATTGGGCGCGAGGCCGTGG | Dnajc2 | NM_009584 | No | 10123.5 | 921 | 258 | Effective | 0.620607387 | 321 | No | 0.246205597 | 8 | 11 | It is not a hit | No |
| 191 | CCGAAGAGCTCTCCCTTGGG | Adnp | NM_009628 | No | 6537.5 | 315 | 0 | Not effective | 1.545809176 | 58 | No | -0.002411042 | 55 | 35 | It is not a hit | No |
| 192 | GGACTCATTGTCCAGCTGGG | Adnp | NM_009628 | No | 4461 | 345 | 0 | Not effective | 0.427844261 | 361 | No | -6.80E-04 | 12 | 14 | It is not a hit | No |
| 193 | ACAAGATCCAGACAGCTCAG | Brca1 | NM_009764 | No | 3092.5 | 135 | 2 | Not effective | 0.95098277 | 220 | No | 0.024155583 | 1 | 21 | It is not a hit | No |
| 194 | ACGGAAGAGAAAGAATTTCG | Brca1 | NM_009764 | No | 1206.5 | 65 | 0 | Not effective | 0.138080942 | 394 | No | -0.040292726 | 0 | 17 | It is not a hit | No |
| 195 | AGCTCGGCTGTTTCCGTGAG | Creb1 | NM_009952 | No | 5614 | 259 | 0 | Not effective | 0.534143096 | 343 | No | 0.024148119 | 417 | 25 | It is not a hit | No |
| 196 | GAAGGCCTTCGGCTACTTCC | Creb1 | NM_009952 | No | 8092 | 548 | 0 | Not effective | 1.552357494 | 56 | No | -0.007011422 | 1 | 31 | It is not a hit | No |

Appendix C: Related to pooled sgRNA library, CRISPRa ssRNA-seq dataset with three replicates merged, screen results and information of sgRNAs used for pilot and validation experiments.

| sgRNA ID | sgRNA target sequence | Target gene | Target transcript | Off-targets | Number of total mapped reads in pilot plasmid library | Number of cells expressing the sgRNA in Genomics pooled CRISPRa dataset | Number of cells with transduction (FDR 10%, chapter 4) | Effective/not effective (more than 10 cells activated) classification in chapter 4 | Fraction of variance explained by NMF4 factor 3 | Hit rank | Screen hit (Yes/No) | Mean log fold change expression of target gene in 10X Genomics pooled screen CRISPRa dataset | Number of DE genes (FDR 10%) in 10X Genomics pooled screen CRISPRa dataset | Number of genes recovered from 214 sgRNAs (from Appendix I) in 10X Genomics pooled screen CRISPRa dataset | Hit above background in differential gene expression rank | sgRNA used in other experiments |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 197 | GGAGCGCGATTTAAAGGTGC | Glo2 | NM_030262 | No | 151 | 146 | 4 | Not effective | 1.10063804 | 146 | No | 0.010972729 | 0 | 30 | No | No |
| 198 | TGGCGTACGCGCCGCCGGGT | Glo2 | NM_030262 | No | 168 | 152 | 17 | Not effective | 0.47431379 | 152 | No | 0.027333851 | 0 | 31 | No | No |
| 199 | AGCCGCGCGCGCCGCCGGGT | Glo2 | NM_030262 | No | 168 | 152 | 17 | Not effective | 0.47431379 | 152 | No | 0.027333851 | 0 | 31 | No | No |
| 200 | GGCTCACTCGCGAGCCGTTG | Hnf4a | NM_030710 | No | 531 | 300 | 47 | Not effective | 0.70496043 | 300 | No | 0.078497045 | 0 | 26 | No | No |
| 201 | GAGCTTTCCCTTCCGATTCC | Smad2 | NM_030754 | No | 8628.5 | 4570 | 724 | Not effective | 0.70496043 | 300 | No | -0.045709769 | 0 | 18 | No | No |
| 202 | GGGTGTTGATGGCAGCCGTG | Smad2 | NM_030754 | No | 8628.5 | 4570 | 724 | Not effective | 0.70496043 | 300 | No | -0.00354725 | 0 | 22 | No | No |
| 203 | TGGCGATCCGCAGCCGCTGT | Mdm4 | NM_030869 | No | 8981.5 | 4570 | 724 | Not effective | 1.19541893 | 1564 | No | 0.013328761 | 0 | 19 | No | No |
| 204 | TGGCGATCCGCAGCCGCTGT | Mdm4 | NM_030869 | No | 8981.5 | 4570 | 724 | Not effective | 1.19541893 | 1564 | No | 0.013328761 | 0 | 19 | No | No |
| 205 | GATAGGTGGCTGAGAAGCGG | Rhbdd3 | NM_030901 | No | 4071 | 310 | 43 | Not effective | 0.07819521 | 1398 | No | -0.007353021 | 0 | 24 | No | No |
| 206 | GCGACGAGCCCTGCCTCCCT | Rhbdd3 | NM_030901 | No | 4071 | 310 | 43 | Not effective | 0.07819521 | 1398 | No | -0.007353021 | 0 | 24 | No | No |
| 207 | GAGGGCGGCTCTGCTAGCCG | Mfsd1 | NM_030913 | No | 5467.5 | 452 | 62 | Not effective | 0.92970149 | 2330 | No | -0.011965987 | 0 | 17 | No | No |
| 208 | AGCGTTCGCGCTGCAGCTGG | Mfsd1 | NM_030913 | No | 5467.5 | 452 | 62 | Not effective | 0.92970149 | 2330 | No | -0.011965987 | 0 | 17 | No | No |
| 209 | GGCTTCGCGCTGCAGCTGGC | Mfsd1 | NM_030913 | No | 5467.5 | 452 | 62 | Not effective | 0.92970149 | 2330 | No | -0.011965987 | 0 | 17 | No | No |
| 210 | CTATCAGCTTTAGTTGCTGTG | Uhrf1 | NM_030914 | No | 9383.5 | 155 | 65 | Not effective | 0.50503749 | 3132 | No | 0.256181132 | 0 | 25 | No | No |
| 211 | GGCTCTGATGTTGCTCCTTC | Uhrf1 | NM_030914 | No | 9383.5 | 155 | 65 | Not effective | 0.50503749 | 3132 | No | 0.256181132 | 0 | 25 | No | No |
| 212 | GGCTCTGATGTTGCTCCTTC | Uhrf1 | NM_030914 | No | 9383.5 | 155 | 65 | Not effective | 0.50503749 | 3132 | No | 0.256181132 | 0 | 25 | No | No |
| 213 | CTATCAGCTTTAGTTGCTGTG | Uhrf1 | NM_030914 | No | 9383.5 | 155 | 65 | Not effective | 0.50503749 | 3132 | No | 0.256181132 | 0 | 25 | No | No |
| 214 | GGGCGGCGGCGAGCAGGCTG | Rnf1 | NM_030931 | No | 1598 | 121 | 31 | Not effective | 0.69954741 | 731 | No | -0.03028986 | 0 | 6 | No | No |
| 215 | GGGGGGGGCGTGAGAGTGGG | Rnf1 | NM_030931 | No | 1598 | 121 | 31 | Not effective | 0.69954741 | 731 | No | -0.03028986 | 0 | 6 | No | No |
| 216 | TTTGTGTGGCGCGGGCGGCC | Rnf2 | NM_030938 | No | 14596.5 | 938 | 746 | Not effective | 0.89038785 | 316 | No | -0.010304239 | 0 | 5 | No | No |
| 217 | GGGGGGGGCGTGAGAGTGGG | Rnf2 | NM_030938 | No | 14596.5 | 938 | 746 | Not effective | 0.89038785 | 316 | No | -0.010304239 | 0 | 5 | No | No |
| 218 | GGCGGCGGCGAGCAGGCTGG | Rnf2 | NM_030938 | No | 14596.5 | 938 | 746 | Not effective | 0.89038785 | 316 | No | -0.010304239 | 0 | 5 | No | No |
| 219 | GAGCGCGCGCGCGCGGGTGG | Rnf2 | NM_030938 | No | 14596.5 | 938 | 746 | Not effective | 0.89038785 | 316 | No | -0.010304239 | 0 | 5 | No | No |
| 220 | GGCGCGCGGCGAGCAGGCTG | Rnf2 | NM_030938 | No | 14596.5 | 938 | 746 | Not effective | 0.89038785 | 316 | No | -0.010304239 | 0 | 5 | No | No |
| 221 | GGTGCATTGGCGCGCCGTGG | Rnf11 | NM_031047 | No | 4283 | 677 | 177 | Not effective | 1.14879595 | 157 | No | 0.000880789 | 0 | 1 | No | No |
| 222 | GGTGCTTCGGCGCGCCGTGG | Rnf11 | NM_031047 | No | 4283 | 677 | 177 | Not effective | 1.14879595 | 157 | No | 0.000880789 | 0 | 1 | No | No |
| 223 | GGTGCTTCGGCGCGCCGTGG | Rnf11 | NM_031047 | No | 4283 | 677 | 177 | Not effective | 1.14879595 | 157 | No | 0.000880789 | 0 | 1 | No | No |
| 224 | GGCTTCGCTCGCCGCCTCCC | Smarca2 | NM_031116 | No | 7245 | 746 | 199 | Not effective | 1.52131518 | 60 | No | -0.007354011 | 0 | 5 | No | No |
| 225 | CTGGCTCGGCGGCGGAGAGC | Smarca2 | NM_031116 | No | 7245 | 746 | 199 | Not effective | 1.52131518 | 60 | No | -0.007354011 | 0 | 5 | No | No |
| 226 | GCGGCGGAGGATGAGCAAAA | Smarca2 | NM_031116 | No | 7245 | 746 | 199 | Not effective | 1.52131518 | 60 | No | -0.007354011 | 0 | 5 | No | No |
| 227 | GAGTGATTGGGAGCTGCAGT | Tnks | NM_031355 | No | 1896 | 134 | 69 | Not effective | 0.09050747 | 1 | No | -0.018772815 | 0 | 7 | No | No |
| 228 | GCGAGGGAGGATGAGCAAAA | Tnks | NM_031355 | No | 1896 | 134 | 69 | Not effective | 0.09050747 | 1 | No | -0.018772815 | 0 | 7 | No | No |
| 229 | GCGAGGGAGGATGAGCAAAC | Tnks | NM_031355 | No | 1896 | 134 | 69 | Not effective | 0.09050747 | 1 | No | -0.018772815 | 0 | 7 | No | No |
| 230 | GCGCGCGCGCGCGCGCGCGT | Tnks | NM_031355 | No | 1896 | 134 | 69 | Not effective | 0.09050747 | 1 | No | -0.018772815 | 0 | 7 | No | No |
| 231 | CTGGCGCGGCGGCGAGCGAA | Tnks2 | NM_031588 | No | 10045 | 1511 | 47 | Not effective | 0.13784427 | 432 | No | -0.021796782 | 0 | 17 | No | No |
| 232 | CTGGCGCGGCGGCGAGCGAA | Tnks2 | NM_031588 | No | 10045 | 1511 | 47 | Not effective | 0.13784427 | 432 | No | -0.021796782 | 0 | 17 | No | No |
| 233 | GCGCGCGCGCGCGCGCGCGT | Tnks2 | NM_031588 | No | 10045 | 1511 | 47 | Not effective | 0.13784427 | 432 | No | -0.021796782 | 0 | 17 | No | No |
| 234 | CATGATGCGTGAGGGTTTTC | Tnfrsf1a | NM_031857 | No | 11102.5 | 701 | 10 | Not effective | 0.13295989 | 129 | No | 0.013052062 | 0 | 35 | No | No |
| 235 | TGCGCGCTCGCGACGACTCG | Tnfrsf1a | NM_031857 | No | 11102.5 | 701 | 10 | Not effective | 0.13295989 | 129 | No | 0.013052062 | 0 | 35 | No | No |
| 236 | GCTCGTCGCCGCCGCCGCCG | Tnfrsf1a | NM_031857 | No | 11102.5 | 701 | 10 | Not effective | 0.13295989 | 129 | No | 0.013052062 | 0 | 35 | No | No |
| 237 | ATGTATGATTTGCGTCCTAT | Ern1 | NM_031934 | No | 5715.5 | 1531 | 199 | Not effective | 1.17741991 | 36 | Yes | -0.055847383 | 0 | 41 | No | No |
| 238 | GCCTATTGCTCACGCCTCCC | Ern1 | NM_031934 | No | 5715.5 | 1531 | 199 | Not effective | 1.17741991 | 36 | Yes | -0.055847383 | 0 | 41 | No | No |
| 239 | TTAAAATGCAGAATGTGAGC | Ern1 | NM_031934 | No | 5715.5 | 1531 | 199 | Not effective | 1.17741991 | 36 | Yes | -0.055847383 | 0 | 41 | No | No |
| 240 | GCTCTCGCCAGCGCCGTCAT | Mdm2 | NM_031995 | No | 8973 | 1595 | 90 | Not effective | 0.18718084 | 1102 | No | 0.001581102 | 0 | 26 | No | No |
| 241 | ATCTGCTGTGCTCTTCTGGA | Mdm2 | NM_031995 | No | 8973 | 1595 | 90 | Not effective | 0.18718084 | 1102 | No | 0.001581102 | 0 | 26 | No | No |
| 242 | GTGTGGAGAGGACAAGCGTT | Phf20l1 | NM_032030 | No | 4579 | 204 | 79 | Not effective | 0.32029047 | 396 | No | -0.01396499 | 0 | 17 | No | No |
| 243 | GGGCTGGCTGCTGCATAGAT | Mgs | NM_032720 | No | 14062 | 1504 | 62 | Not effective | 0.65934824 | 1630 | No | -0.009684835 | 0 | 19 | No | No |
| 244 | AGAAATGGCACTGGGCATCG | Mgs | NM_032720 | No | 14062 | 1504 | 62 | Not effective | 0.65934824 | 1630 | No | -0.009684835 | 0 | 19 | No | No |
| 245 | GCGCTGGCTGCTGCATAGAT | Phf20 | NM_032783 | No | 7202 | 1504 | 62 | Not effective | 0.65934824 | 1630 | No | -0.009684835 | 0 | 19 | No | No |
| 246 | ATGAGAGTCTGTCATTGCTG | Mdm5 | NM_033758 | No | 5673 | 1188 | 82 | Not effective | 0.44758167 | 3358 | No | 0.004481457 | 0 | 21 | No | No |
| 247 | GAATTAAGTCGTCTTGTCGC | Mdm5 | NM_033758 | No | 5673 | 1188 | 82 | Not effective | 0.44758167 | 3358 | No | 0.004481457 | 0 | 21 | No | No |
| 248 | GCGTCACTTCTGTCTTCGGA | Mdm5 | NM_033758 | No | 5673 | 1188 | 82 | Not effective | 0.44758167 | 3358 | No | 0.004481457 | 0 | 21 | No | No |
| 249 | TCTGCGTAAGCGGCGCGCTC | Smad5 | NM_033759 | No | 8980.5 | 1620 | 371 | Not effective | 0.07257746 | 439 | No | -0.008716527 | 0 | 21 | No | No |
| 250 | TCTGCGCAGCAAAGCCGTCC | Smad5 | NM_033759 | No | 8980.5 | 1620 | 371 | Not effective | 0.07257746 | 439 | No | -0.008716527 | 0 | 21 | No | No |
| 251 | TCGCTCGCTCAGGGCCCCG | Thx2 | NM_034075 | No | 13416 | 1504 | 624 | Not effective | 0.13695884 | 1111 | No | 0.013572706 | 0 | 17 | No | No |
| 252 | TCGCTCGCTCAGGGCCCCG | Thx2 | NM_034075 | No | 13416 | 1504 | 624 | Not effective | 0.13695884 | 1111 | No | 0.013572706 | 0 | 17 | No | No |
| 253 | TTTAGAAAGGCGAAGTGGTG | Rnf6 | NM_032795 | No | 1586 | 95 | 5 | Not effective | 0.22295446 | 382 | No | -0.012339595 | 0 | 26 | No | No |
| 254 | ATCACCGATCAATTTATGACC | Rnf6 | NM_032795 | No | 1586 | 95 | 5 | Not effective | 0.22295446 | 382 | No | -0.012339595 | 0 | 26 | No | No |
| 255 | ATTGAGAAGGCGTAAGGTTG | Smarca1 | NM_033817 | No | 1482 | 148 | 0 | Not effective | 0 | 439 | No | -0.012339595 | 0 | 26 | No | No |
| 256 | GCCCCTCTCATCCCCGGGGC | Smarca1 | NM_033817 | No | 1482 | 148 | 0 | Not effective | 0 | 439 | No | -0.012339595 | 0 | 26 | No | No |
| 257 | TCCCCCCTCTCATCCCCGGG | Nfx1 | NM_033823 | No | 9053.5 | 1304 | 82 | Not effective | 0.61282008 | 1320 | No | 0.009833538 | 0 | 32 | No | No |
| 258 | AGAGTATGGTTGGTAGGAAT | Snx6a1 | NM_033877 | No | 7223 | 1139 | 2 | Not effective | 0.11397071 | 1581 | No | 0.013549644 | 0 | 11 | No | No |
| 259 | GGAATAAATTTGGTCCCTTC | Snx6a1 | NM_033877 | No | 7223 | 1139 | 2 | Not effective | 0.11397071 | 1581 | No | 0.013549644 | 0 | 11 | No | No |
| 260 | GCGCGCAGAGCGCAAGCAGC | Arid5b | NM_033889 | No | 3426.5 | 188 | 0 | Not effective | 0.04343772 | 424 | No | -0.000343786 | 0 | 14 | No | No |
| 261 | AACCGTATGGGTTGCGGAGG | Arid5b | NM_033889 | No | 3426.5 | 188 | 0 | Not effective | 0.04343772 | 424 | No | -0.000343786 | 0 | 14 | No | No |
| 262 | GGAAGCAGAGCAGCCGTCC | Fam3 | NM_033760 | No | 10154 | 894 | 5 | Not effective | 0.13827636 | 1176 | No | -0.013736171 | 0 | 32 | No | No |
| 263 | TGTGCGCCGCCGCCTCCG | Fam3 | NM_033760 | No | 10154 | 894 | 5 | Not effective | 0.13827636 | 1176 | No | -0.013736171 | 0 | 32 | No | No |
| 264 | CTGCAGCTCCGCCACCCGTT | Sod1 | NM_033776 | No | 9914 | 766 | 7 | Not effective | 0.13819965 | 1378 | No | -0.009813094 | 0 | 29 | No | No |
| 265 | GCGCGCGCGCGCGCGCGCGT | Sod1 | NM_033776 | No | 9914 | 766 | 7 | Not effective | 0.13819965 | 1378 | No | -0.009813094 | 0 | 29 | No | No |
| 266 | ATCGTAGAAGAGAGTGCCGT | Bcl6 | NM_033809 | No | 10011 | 1116 | 0 | Not effective | 0.07738787 | 235 | No | -0.009837611 | 0 | 21 | No | No |
| 267 | ATGCGCGCGGCAGCGAGTCC | Bcl6 | NM_033809 | No | 10011 | 459 | 0 | Not effective | 0.07738787 | 235 | No | -0.009837611 | 0 | 21 | No | No |
| 268 | GGCGCGCGTCGAGCGCAGTG | Mphosph8 | NM_033773 | No | 12748 | 653 | 171 | Not effective | 1.19479267 | 186 | No | -0.009834721 | 0 | 18 | No | No |
| 269 | GAACTCTTTAAAGCAGCAGG | Mphosph8 | NM_033773 | No | 12748 | 653 | 171 | Not effective | 1.19479267 | 186 | No | -0.009834721 | 0 | 18 | No | No |
| 270 | CTCCTCCCTCGCAGCGCGG | Mgea5 | NM_033799 | No | 9780.5 | 784 | 0 | Not effective | 1.14860108 | 411 | No | -0.00374227 | 0 | 18 | No | No |
| 271 | TGAACTGGCCCTCCTGGGAA | Phf10 | NM_034250 | Multi-mapping | 15742.5 | 1124 | 0 | Not effective | 0.72290923 | 282 | No | -0.007725455 | 0 | 25 | No | No |
| 272 | ATCCCCCTCCCTCGCAGCGC | Phf10 | NM_034250 | Multi-mapping | 15742.5 | 1124 | 0 | Not effective | 0.72290923 | 282 | No | -0.007725455 | 0 | 25 | No | No |
| 273 | GGGCGCAGCGCAAGCAGCAC | Cum1 | NM_031531 | No | 10658 | 517 | 0 | Not effective | 0.78918897 | 275 | No | -0.009684835 | 0 | 19 | No | No |
| 274 | CTCCCTCAGCGCCGCAGAGG | Cum1 | NM_031531 | No | 10658 | 517 | 0 | Not effective | 0.78918897 | 275 | No | -0.009684835 | 0 | 19 | No | No |
| 275 | TCTCCACCTTGCAGCCCAGA | Hgf | NM_032826 | No | 9981.5 | 141 | 0 | Not effective | 0.13828845 | 1115 | No | -0.009834721 | 0 | 21 | No | No |
| 276 | TCTCCACCTTGCAGCCCAGA | Hgf | NM_032826 | No | 9981.5 | 141 | 0 | Not effective | 0.13828845 | 1115 | No | -0.009834721 | 0 | 21 | No | No |
| 277 | GGGCGCGCGTCGAGCGCAGTG | Mphosph8 | NM_033773 | No | 12748 | 653 | 171 | Not effective | 1.19479267 | 186 | No | -0.009834721 | 0 | 18 | No | No |
| 278 | GGGCGCGCGTCGAGCGCAGTG | Mgea5 | NM_033799 | No | 9780.5 | 784 | 0 | Not effective | 1.14860108 | 411 | No | -0.00374227 | 0 | 18 | No | No |
| 279 | CTCCTCCCTCGCAGCGCGGG | Phf10 | NM_034250 | Multi-mapping | 15742.5 | 1124 | 0 | Not effective | 0.72290923 | 282 | No | -0.007725455 | 0 | 25 | No | No |
| 280 | CAGCGCGCTGGCGGCCATGG | Phf10 | NM_034250 | Multi-mapping | 15742.5 | 1124 | 0 | Not effective | 0.72290923 | 282 | No | -0.007725455 | 0 | 25 | No | No |
| 281 | GTCATGAATAGTGAGCGCCG | Irf3 | NM_032854 | No | 8096.5 | 1029 | 153 | Not effective | 0.08898219 | 241 | No | -0.00773551 | 0 | 17 | No | No |
| 282 | TGCCCTCGAGGAAGAGCGGA | Irf3 | NM_032854 | No | 8096.5 | 1029 | 153 | Not effective | 0.08898219 | 241 | No | -0.00773551 | 0 | 17 | No | No |
| 283 | TCGTGTAAACGTGTAGAGGA | Rhg20a | NM_032812 | No | 2378 | 96 | 92 | Not effective | 0.04507089 | 428 | No | -0.002814427 | 0 | 15 | No | No |
| 284 | TCGTGTAAACGTGTAGAGGA | Rhg20a | NM_032812 | No | 2378 | 96 | 92 | Not effective | 0.04507089 | 428 | No | -0.002814427 | 0 | 15 | No | No |
| 285 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 286 | TCGAGCCGCCTCCCCCACTCC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 287 | GAGCGCCTCTCACTCGTAGGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 288 | TTCTCCTTCTCATTGCTTGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 289 | TAAGTCCTTGCGGTCACTTC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 290 | GCGTGTGACAGCAGTGTCCC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 291 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 292 | GAGCGCCTCTCACTCGTAGGC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 293 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 294 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 295 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 296 | CCGCGCACTTACTCATTCCA | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 297 | GAGCGAAGCGACGCTCCACTC | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 298 | CCGCGCACTTACTCATTCCA | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 299 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 300 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 301 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 302 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 303 | GAGCGCCTCTCACTCGTAGGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 304 | TTCTCCTTCTCATTGCTTGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 305 | TAAGTCCTTGCGGTCACTTC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 306 | GCGTGTGACAGCAGTGTCCC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 307 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 308 | GAGCGCCTCTCACTCGTAGGC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 309 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 310 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 311 | GCGCGCGCGCGCGCGCGCGT | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 312 | GAGCGCCTCTCACTCGTAGGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 313 | TTCTCCTTCTCATTGCTTGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 314 | TAAGTCCTTGCGGTCACTTC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 315 | GCGTGTGACAGCAGTGTCCC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 316 | CCGCGCACTTACTCATTCCA | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 317 | GAGCGAAGCGACGCTCCACTC | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 318 | CCGCGCACTTACTCATTCCA | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 319 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 320 | GAGCGCCTCTCACTCGTAGGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 321 | TTCTCCTTCTCATTGCTTGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 322 | TAAGTCCTTGCGGTCACTTC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 323 | GCGTGTGACAGCAGTGTCCC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 324 | CCGCGCACTTACTCATTCCA | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 325 | GAGCGAAGCGACGCTCCACTC | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 326 | CCGCGCACTTACTCATTCCA | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 327 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 328 | GAGCGCCTCTCACTCGTAGGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 329 | TTCTCCTTCTCATTGCTTGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 330 | TAAGTCCTTGCGGTCACTTC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 331 | GCGTGTGACAGCAGTGTCCC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 332 | CCGCGCACTTACTCATTCCA | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 333 | GAGCGAAGCGACGCTCCACTC | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 334 | CCGCGCACTTACTCATTCCA | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 335 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 336 | GAGCGCCTCTCACTCGTAGGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 337 | TTCTCCTTCTCATTGCTTGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 338 | TAAGTCCTTGCGGTCACTTC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 339 | GCGTGTGACAGCAGTGTCCC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 340 | CCGCGCACTTACTCATTCCA | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 341 | GAGCGAAGCGACGCTCCACTC | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 342 | CCGCGCACTTACTCATTCCA | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 343 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 344 | GAGCGCCTCTCACTCGTAGGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 345 | TTCTCCTTCTCATTGCTTGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 346 | TAAGTCCTTGCGGTCACTTC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 347 | GCGTGTGACAGCAGTGTCCC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 348 | CCGCGCACTTACTCATTCCA | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 349 | GAGCGAAGCGACGCTCCACTC | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 350 | CCGCGCACTTACTCATTCCA | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 351 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 15818 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 352 | GAGCGCCTCTCACTCGTAGGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 353 | TTCTCCTTCTCATTGCTTGC | Mcts1 | NM_032910 | No | 3981 | 130 | 12 | Not effective | 0.07825969 | 203 | No | 0.003273241 | 0 | 20 | No | No |
| 354 | TAAGTCCTTGCGGTCACTTC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 355 | GCGTGTGACAGCAGTGTCCC | Rhd1 | NM_032815 | No | 4710 | 208 | 0 | Not effective | 1.08031379 | 14 | Yes | -0.004158059 | 0 | 25 | No | No |
| 356 | CCGCGCACTTACTCATTCCA | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 357 | GAGCGAAGCGACGCTCCACTC | Rnnt | NM_032840 | No | 642 | 133 | 0 | Not effective | 0.09824303 | 1581 | No | 0.009114479 | 0 | 15 | No | No |
| 358 | CCGCGCACTTACTCATTCCA | Cactbp1 | NM_032821 | No | 14632 | 1663 | 158 | Not effective | 0.78396794 | 286 | No | -0.004313342 | 0 | 22 | No | No |
| 359 | GCGTGTGACAGCAGTGTCCC | Cactbp1 | NM_032821 | No | 15818 | 1663 

Appendix C: Related to pooled sgRNA library, CRISPRa scRNA-seq dataset with three replicates merged, screen results and information of sgRNAs used for pilot and validation experiments

| sgRNA ID | sgRNA target sequence | Target gene | Target transcript | Off-targets | Number of counts average read 1 and read 2 plasmid library | Number of cells expressing the sgRNA in 10X Genomics pooled CRISPRa dataset | Number of cells with activation (FDR 10%, chapter 4) | Effective/not effective (more than 10 cells activated) classification in chapter 4 | Fraction of variance explained by MOFA factor 3 | Hit rank | Screen hit (Yes/No) | Mean log fold change expression of target gene in 10X Genomics pooled screen CRISPRa dataset | Number of DE genes (FDR 10%) in 10X Genomics pooled screen CRISPRa dataset | Number of genes recovered from ZGA signature (from Appendix I) in 10X Genomics pooled screen CRISPRa dataset | Hit above background in differential gene expression rank | sgRNA used in other experiments |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 298 | GCACGGCGCTCAGGGAGGG | Smyd2 | NM_026796 | No | 9732 | 270 | 0 | Not effective | 1.228373697 | 134 | No | -0.078485868 | 75 | 34 | It is not a hit | No |
| 299 | AATTAGACGATGGAGGAACA | Spata3 | NM_027029 | No | 934 | 60 | 2 | Not effective | 0.105188508 | 405 | No | 0.0271732 | 2 | 14 | It is not a hit | No |
| 300 | GTCTTTGGGCCTAAGAGAAC | Spata3 | NM_027029 | No | 1844 | 112 | 0 | Not effective | 2.197197611 | 13 | Yes | -3.82E-04 | 0 | 38 | Yes | No |
| 301 | CAGGTCCTGTGAGCATGCGC | Smyd3 | NM_027188 | No | 11192 | 678 | 13 | Effective | 0.868572371 | 255 | No | 0.050480221 | 2 | 18 | It is not a hit | No |
| 302 | TGTAGCGAAGCGCAGGAGCC | Smyd3 | NM_027188 | No | 5295 | 322 | 2 | Not effective | 0.624934739 | 318 | No | 0.044906242 | 1 | 17 | It is not a hit | No |
| 303 | GGTGGGAGAAGTGCTTGTGG | Zmynd8 | NM_027230 | No | 3599 | 285 | 0 | Not effective | 0.841891128 | 282 | No | 0.065994381 | 1 | 14 | It is not a hit | No |
| 304 | TTTGCTGGAGCAGGTCAGGC | Zmynd8 | NM_027230 | No | 5329 | 359 | 0 | Not effective | 0.996808722 | 201 | No | -0.062879566 | 0 | 30 | It is not a hit | No |
| 305 | ACCGGAGCCGACGGGAACCG | Atad2 | NM_027435 | No | 15487.5 | 865 | 371 | Effective | 1.415298178 | 81 | No | 0.487330129 | 17 | 34 | It is not a hit | No |
| 306 | CTCCTCCAATCCGAAGAGCG | Atad2 | NM_027435 | No | 5476.5 | 316 | 0 | Not effective | 1.839501959 | 29 | Yes | -0.080599735 | 0 | 31 | Yes | No |
| 307 | GCCGCCTCGGGCGGGATTG | Prdm16 | NM_027504 | No | 18109.5 | 1226 | 0 | Not effective | 1.023088545 | 192 | No | -0.007845352 | 128 | 25 | It is not a hit | No |
| 308 | GGTGAAGACCGAGAAGGCCT | Prdm16 | NM_027504 | No | 4088 | 183 | 0 | Not effective | 0.603910492 | 326 | No | 0.012964073 | 7 | 18 | It is not a hit | No |
| 309 | CTGCTGGTACTGCCTCGCCC | Taf3 | NM_027748 | No | 16118 | 1136 | 6 | Not effective | 0.497803124 | 348 | No | 0.001974689 | 5 | 20 | It is not a hit | No |
| 310 | GGCGGAGAGATAGCGGGCCC | Taf3 | NM_027748 | No | 7704 | 438 | 0 | Not effective | 1.85172676 | 27 | Yes | -0.078315271 | 6 | 41 | Yes | No |
| 311 | ACGTTTCCGAGTGAGACAGC | Ints12 | NM_027927 | No | 3898.5 | 240 | 0 | Not effective | 1.283786929 | 117 | No | -0.064880209 | 0 | 24 | It is not a hit | No |
| 312 | GAAACAAAGGTTCCCGCAAC | Ints12 | NM_027927 | No | 4462 | 229 | 22 | Effective | 1.312763067 | 110 | No | 0.170246803 | 4 | 23 | It is not a hit | No |
| 313 | CCAGGGTGCTGATCAGCCAT | Setd3 | NM_028262 | No | 5455 | 333 | 3 | Not effective | 1.805472353 | 33 | Yes | -0.03811091 | 1 | 52 | Yes | No |
| 314 | CTTCGGATACGGCTCCGCCC | Setd3 | NM_028262 | No | 12926.5 | 979 | 4 | Not effective | 1.315789016 | 108 | No | -0.043408627 | 14 | 31 | It is not a hit | No |
| 315 | CCACTCTCTCGACTGCCAGA | Zfp655 | NM_028298 | No | 8906.5 | 623 | 0 | Not effective | 0.336225622 | 368 | No | -0.021019957 | 1 | 15 | It is not a hit | No |
| 316 | GGCGCAGGAGTCCACGCGGA | Zfp655 | NM_028298 | No | 6675.5 | 276 | 2 | Not effective | 0.324851406 | 372 | No | 0.084441128 | 36 | 25 | It is not a hit | No |
| 317 | CTGCGGCGGAGGAGCGCGAT | Tdrkh | NM_028307 | No | 7796 | 398 | 2 | Not effective | 0.861474989 | 259 | No | 0.014374449 | 1 | 19 | It is not a hit | No |
| 318 | TTGGTGAGGATGCGCTGCGG | Tdrkh | NM_028307 | No | 6747 | 423 | 0 | Not effective | 1.234075009 | 130 | No | -0.022563418 | 0 | 14 | It is not a hit | No |
| 319 | CTCCAAGTGGAGTATCCCTG | Tmt11 | NM_028604 | No | 4154 | 77 | 0 | Not effective | 0.123633761 | 397 | No | 0.022761667 | 0 | 17 | It is not a hit | No |
| 320 | TAAGGGCAAGAGTGAGATCC | Tmt11 | NM_028604 | No | 2578 | 155 | 2 | Not effective | 2.115000707 | 14 | Yes | 0.103992235 | 4 | 27 | It is not a hit | No |
| 321 (or Dppa4 sgRNA 1) | GTTGAGGGTGGGACCAGAAG | Dppa4 | NM_028610 | No | 2402.5 | 77 | 13 | Effective | 0 | 436 | No | 0.240578787 | 1 | 22 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 322 (or Dppa4 sgRNA 2) | ATCACAATTTTGTTGAGGGT | Dppa4 | NM_028610 | No | 1696.5 | 108 | 0 | Not effective | 0.116458901 | 400 | No | 0.008375177 | 0 | 20 | It is not a hit | No |
| 323 | ACACAGGAGGACCCTCCCTC | Dppa2 | NM_028615 | No | 10653 | 331 | 170 | Effective | 2.08541854 | 15 | Yes | 0.499990436 | 20 | 50 | Yes | Validated by individual CRISPRa + bulk RNA-seq |
| 324 | CCGCGGCAGAATAGCAGCC | Dppa2 | NM_028615 | No | 4205.5 | 115 | 20 | Effective | 2.64396782 | 5 | Yes | 0.182844676 | 6 | 52 | Yes | No |
| 325 | GGATGTGCACAGTGATTTAG | Elp3 | NM_028811 | No | 1543.5 | 85 | 1 | Not effective | 0.010394968 | 433 | No | 0.079260513 | 0 | 13 | It is not a hit | No |
| 326 | TAGACGCTTCTCCAAGAAGC | Elp3 | NM_028811 | No | 3284.5 | 213 | 0 | Not effective | 0.334282084 | 370 | No | -0.096803394 | 0 | 20 | It is not a hit | No |
| 327 | ACCATACGTTCCGGGAAGCA | Cxxc1 | NM_028868 | No | 4926.5 | 452 | 0 | Not effective | 0.709143462 | 299 | No | -0.025824315 | 4 | 12 | It is not a hit | No |
| 328 | TCGCAACGTCGAAGCAAATG | Cxxc1 | NM_028868 | No | 2975 | 185 | 0 | Not effective | 2.829040057 | 3 | Yes | -0.027801229 | 1 | 30 | It is not a hit | No |
| 329 | GCTGCAGCAGCGACGTGTTA | Tdrd9 | NM_029056 | No | 3969 | 242 | 1 | Not effective | 1.15936954 | 154 | No | 0.01405492 | 4 | 20 | It is not a hit | No |
| 330 | GGCCAGGCCCTGCATGCACG | Tdrd9 | NM_029056 | No | 12459 | 817 | 0 | Not effective | 0.897196091 | 243 | No | 0.00108805 | 1 | 16 | It is not a hit | No |
| 331 | GCTGGCGTAACGATGGGATG | Chd5 | NM_029216 | No | 3216.5 | 248 | 7 | Not effective | 0.429066275 | 360 | No | 0.061334792 | 2 | 17 | It is not a hit | No |
| 332 | GTTAACAGCCGAGAACAAGT | Chd5 | NM_029216 | No | 1532.5 | 74 | 4 | Not effective | 0 | 436 | No | 0.075500273 | 0 | 25 | It is not a hit | No |
| 333 | CGCAAGGCCTTTGGAAGG | Ep400 | NM_029937 | No | 5093 | 192 | 115 | Effective | 1.008832903 | 197 | No | 0.656485232 | 32 | 29 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 334 | GCTGTATGCAGGAGGTGCGT | Ep400 | NM_029937 | No | 5123.5 | 239 | 1 | Not effective | 0.39712072 | 363 | No | -0.067204456 | 0 | 20 | It is not a hit | No |
| 335 | ATCGCGGCGACACGCGAGCG | Cdyl2 | NM_029441 | No | 12345 | 655 | 0 | Not effective | 0.985995061 | 206 | No | -0.072307715 | 0 | 23 | It is not a hit | No |
| 336 | TTTGCAGAATGACAGGCTTG | Cdyl2 | NM_029441 | No | 2111 | 163 | 0 | Not effective | 2.492807434 | 8 | Yes | -0.108475513 | 1 | 23 | It is not a hit | No |
| 337 | CTGCGCTCCGGCCAGCTGGC | Tsc22d4 | NM_029805 | No | 13112 | 1042 | 1 | Not effective | 1.098738048 | 142 | No | -0.03198654 | 0 | 29 | It is not a hit | No |
| 338 | GTGAGCAGGCCCGAGAAGGC | Tsc22d4 | NM_029805 | No | 8393 | 278 | 79 | Effective | 1.800294629 | 35 | Yes | 0.332423132 | 2 | 38 | Yes | No |
| 339 | GGAGCGCAGCCCTGGTCACG | Stk31 | NM_029916 | No | 12464 | 647 | 0 | Not effective | 0.781928855 | 279 | No | 0.001569788 | 0 | 13 | It is not a hit | No |
| 340 | GTGCTGCTGCGCTATAGCGG | Stk31 | NM_029916 | No | 5934.5 | 438 | 7 | Not effective | 0.866273991 | 256 | No | 0.069294701 | 4 | 20 | It is not a hit | No |
| 341 | AAGGGATTTGTAGTCTGTGC | Phf23 | NM_030064 | No | 2015 | 84 | 1 | Not effective | 2.006175147 | 21 | Yes | -0.052914316 | 0 | 27 | It is not a hit | No |
| 342 | CCTCCGGAACTTCAGTTCCC | Phf23 | NM_030064 | No | 8447.5 | 648 | 0 | Not effective | 0.886414511 | 248 | No | -0.105430364 | 5 | 27 | It is not a hit | No |
| 343 | AACTGCCGCTGAGCTTGTC | Brd8 | NM_030147 | No | 5181 | 362 | 0 | Not effective | 0.486635107 | 351 | No | -0.002389604 | 1 | 14 | It is not a hit | No |
| 344 | CTGCGCGGGTGGGTAGAAGC | Brd8 | NM_030147 | No | 3457.5 | 273 | 9 | Not effective | 1.298263672 | 114 | No | 0.105240901 | 2 | 26 | It is not a hit | No |
| 345 | GTGGGTTCCGGCTGCGCGGG | Zeche4 | NM_030185 | No | 20060 | 1027 | 0 | Not effective | 1.230693333 | 132 | No | -0.022589307 | 1 | 27 | It is not a hit | No |
| 346 | TCTCGACTCTGGGTCTGACGG | Zeche4 | NM_030185 | No | 6304 | 387 | 1 | Not effective | 0.224465247 | 383 | No | -0.011382368 | 1 | 15 | It is not a hit | No |
| 347 | CTTTAAGGGAGCTCGGTCTG | Trps1 | NM_032000 | No | 5327.5 | 367 | 22 | Effective | 0.542509437 | 341 | No | 0.140598288 | 4 | 13 | It is not a hit | No |
| 348 | GTGCGGCCGGCCGGTCCCGC | Trps1 | NM_032000 | No | 32014.5 | 1795 | 134 | Effective | 0.78653994 | 276 | No | 0.155688013 | 6 | 19 | It is not a hit | No |
| 349 | ACGCCAATTGGGGGCCATGT | Jmjd6 | NM_033398 | No | 6657 | 409 | 0 | Not effective | 0.82713558 | 265 | No | -0.073827426 | 0 | 22 | It is not a hit | No |
| 350 | GCGTGGCGCAAAGCAGGGAG | Jmjd6 | NM_033398 | No | 3885 | 231 | 13 | Effective | 0.226624639 | 381 | No | 0.113100006 | 157 | 19 | It is not a hit | No |
| 351 | CTGCAGCCCTGCTGCTGCTC | Aff4 | NM_033565 | No | 16189 | 1090 | 122 | Effective | 0.679468242 | 304 | No | 0.222925087 | 1 | 17 | It is not a hit | No |
| 352 | TGGCCGCGGCCCATGTGACC | Aff4 | NM_033565 | No | 10133.5 | 440 | 4 | Not effective | 1.015989653 | 196 | No | 0.062377658 | 1 | 25 | It is not a hit | No |
| 353 | GCGTAGGTAGCTGTCCTGGT | Smarca5 | NM_053124 | No | 6794 | 344 | 210 | Effective | 1.966420154 | 22 | Yes | 0.500609467 | 19 | 58 | Yes | No |
| 354 | TGGAGAAGAGAATCCACTCC | Smarca5 | NM_053124 | No | 3120 | 62 | 30 | Effective | 2.429077875 | 10 | Yes | 0.389477225 | 3 | 31 | Yes | Validated by individual CRISPRa + bulk RNA-seq |
| 355 | CTCAGGGACGCGTGTCCCGA | Baz2a | NM_054078 | No | 9638.5 | 542 | 0 | Not effective | 0.943034908 | 228 | No | -0.019964218 | 1 | 15 | It is not a hit | No |
| 356 | GACGGCGGGGTGGCGTCTCCT | Baz2a | NM_054078 | No | 17217 | 1549 | 0 | Not effective | 0.452253976 | 357 | No | -0.096959337 | 240 | 16 | It is not a hit | No |
| 357 | CGCGGCCTGGCGGCCTCGG | Mta1 | NM_054081 | No | 20979.5 | 1400 | 8 | Not effective | 1.270167854 | 119 | No | -0.031567367 | 1 | 25 | It is not a hit | No |
| 358 | GGGACGGAGGGCCGCGGCG | Mta1 | NM_054081 | No | 14433.5 | 837 | 0 | Not effective | 1.226869118 | 136 | No | 1.22E-05 | 21 | 32 | It is not a hit | No |
| 359 | CCTCATAACTATCTGCCT | Ing4 | NM_133345 | No | 1510.5 | 102 | 0 | Not effective | 0.091980094 | 408 | No | -0.06957758 | 1 | 20 | It is not a hit | No |
| 360 | CGGAAGTTCAGGTACGGAAA | Ing4 | NM_133345 | No | 1768 | 83 | 2 | Not effective | 0.054886188 | 422 | No | -0.004818495 | 0 | 12 | It is not a hit | No |
| 361 | AGTAAAGCATAAACCACAAG | Acaca | NM_133360 | No | 1151 | 58 | 0 | Not effective | 2.941620101 | 2 | Yes | 0.007328928 | 0 | 33 | Yes | No |
| 362 | TCTTGTGTGTTCTCATCTAT | Acaca | NM_133360 | No | 1261.5 | 134 | 0 | Not effective | 0.048059554 | 423 | No | -0.040491898 | 2 | 14 | It is not a hit | No |
| 363 | CCGAGGCGCCGCGCCTTACA | Prmt3 | NM_133740 | No | 15626.5 | 945 | 1 | Not effective | 1.427440556 | 78 | No | 0.082595402 | 15 | 25 | It is not a hit | No |
| 364 | CGACTTATCGCAGCGCCCTC | Prmt3 | NM_133740 | No | 14804 | 843 | 0 | Not effective | 0.974243869 | 212 | No | -0.042715395 | 7 | 23 | It is not a hit | No |
| 365 | GCCCTCCTGCGGAGCGATCT | Ubr | NM_133815 | No | 10757.5 | 609 | 17 | Effective | 1.42895254 | 77 | No | 0.05418923 | 3 | 32 | It is not a hit | No |
| 366 | GGACGCCGGACGCTCACGAG | Ubr | NM_133815 | No | 11293 | 316 | 81 | Effective | 1.024038798 | 191 | No | 0.301255302 | 8 | 26 | It is not a hit | No |
| 367 | ACATTTCCCAGAGTGTTCCG | Nfxl1 | NM_133921 | No | 6341.5 | 317 | 1 | Not effective | 0.78705973 | 274 | No | 0.054645493 | 4 | 15 | It is not a hit | No |
| 368 | GGCGCCGTGGCGCCGTCGGTA | Nfxl1 | NM_133921 | No | 16397 | 644 | 2 | Not effective | 0.641131064 | 313 | No | 0.026334653 | 3 | 28 | It is not a hit | No |
| 369 | GCTACAGCCGCCGAGCGTCG | Psip1 | NM_133948 | No | 11441 | 477 | 0 | Not effective | 1.276965637 | 118 | No | -0.063474341 | 0 | 40 | It is not a hit | No |
| 370 | TGGTACGGCCCAGCGCTAGC | Psip1 | NM_133948 | No | 7704 | 385 | 0 | Not effective | 0.589786475 | 328 | No | -0.08237414 | 0 | 19 | It is not a hit | No |
| 371 | CCGAGGGTCCCAGCCTGGAG | Mbnd1 | NM_134012 | No | 11382.5 | 474 | 10 | Effective | 0.675133536 | 305 | No | 0.120227192 | 616 | 29 | It is not a hit | No |
| 372 | CTAAGTGTTGTGATATCCTG | Mbnd1 | NM_134012 | No | 1679 | 119 | 0 | Not effective | 0.948282147 | 224 | No | 0.028970672 | 1 | 31 | It is not a hit | No |
| 373 | GCTCCCGGAGCTGGTTAGTC | Ybx3 | NM_139117 | No | 7743.5 | 434 | 6 | Not effective | 0.836119848 | 263 | No | 0.048741018 | 1 | 18 | It is not a hit | No |
| 374 | GGCGCCGCGGCTACTCGGCA | Ybx3 | NM_139117 | No | 16399 | 688 | 235 | Effective | 0.89767227 | 242 | No | 0.390421799 | 33 | 33 | It is not a hit | No |
| 375 | AAGACCGGTTCTGGAGAGAA | Satb2 | NM_139146 | No | 1430.5 | 102 | 11 | Effective | 1.007393323 | 198 | No | 0.087028614 | 3 | 26 | It is not a hit | No |
| 376 | GGAAGCGAGGACTCGTCCAA | Satb2 | NM_139146 | No | 4316 | 96 | 8 | Not effective | 2.617975727 | 6 | Yes | 0.085572208 | 31 | 36 | Yes | No |
| 377 | ATTTGGTAAAGAATGGCTTT | Plac8 | NM_139198 | No | 1687.5 | 153 | 121 | Effective | 0.929783648 | 229 | No | 1.243151481 | 6 | 26 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 378 | TTTCCTCCTTACAATTCCAA | Plac8 | NM_139198 | No | 1553.5 | 157 | 102 | Effective | 0 | 436 | No | 1.102807853 | 4 | 18 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 379 (or Dppa3 sgRNA 1) | GAACTGGCTGGGATTGGGCA | Dppa3 | NM_139218 | No | 7910.5 | 377 | 339 | Effective | 1.32251453 | 105 | No | 1.834160615 | 146 | 46 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 380 (or Dppa3 sgRNA 2) | TTAGCATTAAAGTAGCCTGG | Dppa3 | NM_139218 | No | 1613.5 | 84 | 32 | Effective | 0.624736755 | 319 | No | 0.558357528 | 1 | 19 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 381 | GAGAATAAAGTTCCTCATGT | Gatad2b | NM_139304 | No | 1749.5 | 147 | 0 | Not effective | 1.85270253 | 26 | Yes | -0.110980061 | 1 | 36 | Yes | No |
| 382 | GTGAAGTGAAAGAAATGCTG | Gatad2b | NM_139304 | No | 870.5 | 34 | 0 | Not effective | 0 | 436 | No | -0.122693965 | 0 | 16 | It is not a hit | No |
| 383 | CCCAATTTGGTCAGCAGGCG | Cln7 | NM_144811 | No | 7337 | 464 | 0 | Not effective | 1.300242382 | 113 | No | -0.124807774 | 2 | 28 | It is not a hit | No |
| 384 | GTGCCTCTGAAAGCGCCCAG | Cln7 | NM_144811 | No | 10000.5 | 399 | 0 | Not effective | 0.868650833 | 254 | No | -0.088749823 | 10 | 33 | It is not a hit | No |
| 385 | GGTGAATTTAGCGCTTCACC | Ccde1 | NM_144901 | No | 3068.5 | 162 | 0 | Not effective | 0.753324275 | 286 | No | -0.121294006 | 0 | 19 | It is not a hit | No |
| 386 | TAGTGACTCTGCTGCGACGC | Ccde1 | NM_144901 | No | 5901.5 | 344 | 1 | Not effective | 0.494105621 | 349 | No | -0.110141106 | 2 | 15 | It is not a hit | No |
| 387 | GCGATCGCCGCTATGGAGGC | Prmt7 | NM_145404 | No | 7572.5 | 208 | 6 | Not effective | 1.683097569 | 43 | Yes | 0.003682804 | 9 | 33 | Yes | No |
| 388 | TCCATCCAGTTCAGCGGGCG | Prmt7 | NM_145404 | No | 6034.5 | 430 | 2 | Not effective | 1.086539566 | 173 | No | -0.063258942 | 0 | 16 | It is not a hit | No |
| 389 | CGCGTGCAGCCGGGAGGAGG | Ehmt2 | NM_145830 | No | 7980.5 | 200 | 0 | Not effective | 0.432252976 | 359 | No | -0.060872278 | 18 | 31 | It is not a hit | No |
| 390 | GGGCGCAGCGGGCGTTGCAG | Ehmt2 | NM_145830 | No | 10712.5 | 343 | 25 | Effective | 1.095680428 | 169 | No | 0.142506235 | 25 | 27 | It is not a hit | No |
| 391 | CCTCCCTTTCCCTCTGGCTC | Lin28a | NM_145833 | No | 13496 | 644 | 230 | Effective | 1.179980481 | 147 | No | 0.406384797 | 122 | 39 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 392 | GTCAGAGACCAGAGCAGTGG | Lin28a | NM_145833 | No | 3836.5 | 364 | 207 | Effective | 1.018943915 | 194 | No | 0.641458506 | 111 | 13 | It is not a hit | Validated by individual CRISPRa + bulk RNA-seq (Appendix J) |
| 393 | GCGGGGAAATCCCACCGCAC | Chd4 | NM_145979 | No | 15082.5 | 698 | 11 | Effective | 0.907304005 | 239 | No | -0.040318545 | 2 | 14 | It is not a hit | No |
| 394 | TATTATGGGCTGTCCGTGGG | Chd4 | NM_145979 | No | 2875 | 118 | 2 | Not effective | 0.032769436 | 427 | No | -0.020228546 | 5 | 14 | It is not a hit | No |
| 395 | CGCCGCCGAGGAGGAGGAGG | Chd3 | NM_146019 | Multi-mapping | 6231 | 94 | 6 | Not effective | 0.111942113 | 403 | No | 0.055088623 | 17 | 17 | It is not a hit | No |
| 396 | GGTGGTGGTGGTAGCGGTGG | Chd3 | NM_146019 | No | 5282.5 | 194 | 48 | Effective | 0.964523413 | 215 | No | 0.300672031 | 18 | 31 | It is not a hit | No |
| 397 | GCGCACGCCGTCCTGACGCT | Tdrd7 | NM_146142 | No | 19544.5 | 978 | 0 | Not effective | 1.160803834 | 152 | No | -1.53E-04 | 2 | 23 | It is not a hit | No |
| 398 | TCACACGCGGACCTCGAAGT | Tdrd7 | NM_146142 | No | 4847 | 231 | 88 | Effective | 0.472243531 | 353 | No | 0.359977337 | 10 | 27 | It is not a hit | No |

Appendix C. Ranked by pooled sgRNA library, CRISPRa sgRNA-seq dataset with three replicates merged, screen results and information of sgRNAs used for pilot and validation experiments

| sgRNA ID | sgRNA target sequence | Target gene | Target transcript | Off-targets | Number of counts average read 1 and read 2 pooled library | Number of cells expressing the sgRNA CRISPRa element pooled | Number of cells with activation (FDR 10%, chapter 4) | Effected/not effective (more detailed description in chapter 4 | Fraction of variance explained by MAGECK factor 3 | Hit rank | Screen hit (Yes/No) | Mean log fold change expression of target gene pooled screen CRISPRa element | Number of DE genes (FDR 10%) in pooled screen CRISPRa element | Number of genes recovered from ZGA signature (from Appendix J) in SDE Genomics pooled screen CRISPRa dataset | Hit above background in validation gene expression test | sgRNA used in other experiments |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 399 | AGCAGCCTTCATTAGCTCCCG | E2f4 | NM_148952 | no | 8839 | 336 | 0 | Not effective | 1.9078972705 | 33 | Yes | -0.8705827764 | 3 | 26 | No | No |
| 400 | CTCTCGCCGGAAATGCGATATT | Gm15 | NM_153310 | no | 10253 | 243 | 0 | Not effective | 0.0907239505 | 152 | No | 0.0902312579 | 0 | 0 | No | No |
| 401 | CTTCACCCGCGAAATGCGATAAT | Gm15 | NM_153310 | no | 4098.5 | 243 | 0 | Not effective | 0.0907239505 | 154 | No | 0.0902312579 | 0 | 0 | No | No |
| 402 (or Kdm5b sgRNA 1) | AGAGGTTAAGTGATAAGCATTA | Kdm5b | NM_152895 | no | 2745.5 | 14 | 0 | Not effective | 0 | 156 | No | 0.0852838175 | 0 | 0 | No | No |
| 403 (or Kdm5b sgRNA 2) | GAATTGTCTCTTACCAGCGCC | Kdm5b | NM_152895 | no | 5964.5 | 356 | 0 | Not effective | 0.1458159596 | 121 | No | 1.1429215915 | 0 | 0 | No | No |
| 406 | GCAGCAGTAGCTTCAGCCAG | Pou1 | NM_153556 | no | 6900.5 | 234 | 0 | Not effective | 0.1781113075 | 364 | No | 0.0841179401 | 0 | 0 | No | No |
| 407 | CTGGGAAGGGGGAGGGGGTG | Smim4 | NM_172429 | no | 6677.5 | 236 | 0 | Not effective | 1.1607212143 | 45 | Yes | 0.0205779081 | 0 | 0 | No | No |
| 408 | GCAATGGAAGCTGCTTAAA | Smim4 | NM_172429 | no | 6971 | 601 | 10 | Effective | 1.1611838093 | 156 | No | 0.1139410006 | 2 | 2 | No | No |
| 409 | GCAGCTGGAAGCGCCTTTA | Wdr41 | NM_172598 | no | 6914.5 | 225 | 0 | Not effective | 0.9892252507 | 165 | No | -0.0009537821 | 0 | 0 | No | No |
| 410 | GTGGGAGCGGAAAGCGGAAG | Wdr41 | NM_172598 | no | 7183.5 | 225 | 0 | Not effective | 0.8441818417 | 94 | No | -0.0010203133 | 3 | 3 | No | No |
| 411 | ATGACGAAGTCGACTGAACC | Pof13 | NM_177705 | no | 6958.5 | 275 | 0 | Not effective | 1.1370877815 | 111 | No | -0.0010902126 | 0 | 0 | No | No |
| 412 | TACTCCCTACGCTTGCGCCC | Pof13 | NM_177705 | no | 5980 | 275 | 0 | Not effective | 1.1452745796 | 111 | No | -0.0010203133 | 0 | 0 | No | No |
| 413 | CTTCCCACGTTCCACTCCCC | Pde1b | NM_177744 | no | 7057.5 | 1239 | 0 | Not effective | 0.7821488085 | 278 | No | -0.0029607883 | 17 | 17 | No | No |
| 414 | CCGGGAGGGGCGGGAGCTCCGA | Pde1b | NM_177744 | no | 17427 | 1239 | 0 | Not effective | 0.7821488085 | 270 | No | 0.0001040731 | 1 | 1 | No | No |
| 415 | GGTCCGCGGGCGAAGCTCGCGA | Tox3 | NM_172913 | no | 6977.5 | 537 | 0 | Not effective | 0.0897355938 | 272 | No | 0.0029786947 | 1 | 1 | No | No |
| 416 | AGCGCGAGCATCGCTCCTC | Tox3 | NM_172913 | no | 7011.5 | 541 | 5 | Effective | 0.9735748149 | 272 | No | 0.0029786947 | 10 | 10 | No | No |
| 417 | GAGGCTGGCGCGAGGAGCGC | Rbpjl | NM_175111 | no | 10276 | 1350 | 0 | Not effective | 0.7740035 | 158 | No | -0.0010551961 | 2 | 2 | No | No |
| 418 | GAAGCAGGCAGCGCGCAGCCC | Rbpjl | NM_175111 | no | 6688.5 | 447 | 4 | Effective | 1.1775238628 | 32 | Yes | -0.0025533197 | 14 | 14 | No | No |
| 419 | ACGCATCAAATGATCATTC | Avol | NM_175251 | no | 7390 | 490 | 1 | Not effective | 1.2773533007 | 16 | Yes | -0.0059545855 | 5 | 5 | No | No |
| 420 | AGCGATCAAATGATCATTC | Avol | NM_175251 | no | 5187 | 490 | 1 | Not effective | 0.7515333321 | 16 | Yes | -0.0059545855 | 5 | 5 | No | No |
| 421 | TCCTCCTTCGCCGCCTCCTC | Myln1 | NM_177239 | no | 5197.5 | 27 | 4 | Effective | 1.1226747413 | 133 | No | 0.0823110088 | 7 | 7 | No | No |
| 422 | GCGCTTGCGCCGCTTCCTC | Myln1 | NM_177239 | no | 8312 | 31 | 1 | Not effective | 0.8823110088 | 131 | No | 0.0823110088 | 29 | 29 | No | No |
| 423 | GCCTTGCGCTCGCCGCTTC | Myln1 | NM_177239 | no | 11925.5 | 27 | 27 | Effective | 1.1226747413 | 133 | No | 0.0823110088 | 5 | 5 | No | No |
| 424 | GCCTTGCGCTCGCCGCTTC | Myln1 | NM_177239 | no | 13920.5 | 713 | 18 | Effective | 1.1424353708 | 32 | No | 0.0598983448 | 36 | 36 | No | No |
| 425 | GAGAGACATCAGCTCCTC | Ddx60 | NM_177521 | no | 11601 | 1106 | 0 | Not effective | 0.7849828855 | 177 | No | 0.0010137823 | 0 | 0 | No | No |
| 426 | AGATGGCAAGTGCTGCAACGAA | Ddx60 | NM_177521 | no | 4000 | 1106 | 0 | Not effective | 0.7849828855 | 177 | No | 0.0010137823 | 0 | 0 | No | No |
| 427 | GGCGTTGGCGCTGCGCTCCTC | Ddx60 | NM_177521 | no | 7115 | 1106 | 0 | Not effective | 0.7849828855 | 177 | No | 0.0010137823 | 0 | 0 | No | No |
| 428 | GGCGTTGGCGCTGCGCTCCTAAA | Ddx60 | NM_177521 | no | 2216 | 161 | 0 | Not effective | 1.1139657895 | 159 | No | -0.0045563554 | 0 | 0 | No | No |
| 429 | CGGAGAGTCTATTCTCAAA | Rhbg04 | NM_177857 | no | 2023 | 119 | 0 | Not effective | 0.3401178666 | 367 | No | -0.0045563554 | 0 | 0 | No | No |
| 430 | CTTCAAAACTACATTTCCCA | Rhbg04 | NM_177857 | no | 2023 | 119 | 0 | Not effective | 0.3401178666 | 367 | No | -0.0045563554 | 0 | 0 | No | No |
| 431 | CGCCGCCCCTCCTCCAAG | Sirt5 | NM_178750 | no | 10453 | 816 | 0 | Not effective | 0.8779378968 | 159 | No | -0.0005563821 | 0 | 0 | No | No |
| 432 | TGCTCGCCTCTCGCCTCAGG | Sirt5 | NM_178750 | no | 9483 | 504 | 0 | Not effective | 1.1477195158 | 250 | No | -0.0005563821 | 0 | 0 | No | No |
| 433 | ATATAGTTAAAAGATTTCC | Sirt5 | NM_178750 | no | 5385.5 | 504 | 7 | Effective | 1.1477195158 | 107 | No | -0.0005563821 | 21 | 21 | No | No |
| 434 | CACTCTTCTCACTCTCTCTAT | Ccr4 | NM_178848 | no | 77 | 44 | 0 | Not effective | 1.1477533352 | 436 | No | 0.0003550333 | 0 | 0 | No | No |
| 435 | CACTCTTCTCACTCTCTCTAT | Ccr4 | NM_178848 | no | 8737 | 44 | 0 | Not effective | 0 | 436 | No | 0.0003550333 | 0 | 0 | No | No |
| 436 | AAGGGTGAGCTTCCGCAGGC | Mpm2 | NM_181345 | no | 11726.5 | 714 | 1 | Not effective | 0.1167251077 | 158 | No | -0.0010833983 | 0 | 0 | No | No |
| 437 | AAGGGTGAGCTTCCGCAGGC | Mpm2 | NM_181345 | no | 5295.5 | 227 | 5 | Effective | 1.1425823219 | 40 | No | -0.0010833983 | 0 | 0 | No | No |
| 438 | CCTCTCCTTTCCCATTGAAC | Mpm2 | NM_181345 | no | 5455.5 | 189 | 74 | Effective | 0.7232329282 | 32 | No | 0.5533484679 | 0 | 0 | No | No |
| 439 | CCTCTCCTTTCCCATTGAAC | Sirt4 | NM_201396 | no | 5197.5 | 189 | 35 | Effective | 0.7232329282 | 32 | No | 0.5533484679 | 0 | 0 | No | No |
| 440 | CTTCAAAACTACATTTCCCA | Sirt4 | NM_201396 | no | 11215 | 135 | 0 | Not effective | 0.8441818789 | 313 | No | -0.0045563554 | 0 | 0 | No | No |
| 441 | CGCCGCCCCTCCTCCAAG | Sirt4 | NM_201396 | no | 4090.5 | 135 | 0 | Not effective | 0.8441818789 | 313 | No | -0.0045563554 | 0 | 0 | No | No |
| 442 | TGCTCGCCTCTCGCCTCAGG | Tcf12 | NM_207653 | no | 5411 | 601 | 200 | Effective | 0.6580048236 | 307 | No | -0.0005563821 | 1 | 1 | No | No |
| 443 | GGCATCCGCAAAAGAGTAG | Tcf12 | NM_207653 | no | 5449 | 601 | 0 | Not effective | 0.6580048236 | 307 | No | -0.0005563821 | 1 | 1 | No | No |
| 444 | GGCGCCGCGAAAGAGGTAG | Tcf12 | NM_207653 | no | 5455.5 | 601 | 0 | Not effective | 0.6580048236 | 307 | No | -0.0005563821 | 1 | 1 | No | No |
| 445 | GGCGCCGCGAAAGAGGTAGCG | Rasrl | NM_198823 | no | 21000.5 | 1018 | 2 | Not effective | 1.1898184556 | 145 | No | -0.0010833983 | 0 | 0 | No | No |
| 446 | GGGCCCTCTCGCCATCGC | Rasrl | NM_198823 | no | 13559 | 878 | 1 | Not effective | 0.9569048664 | 221 | No | -0.0010509864 | 4 | 4 | No | No |
| 447 | TGAGTGAAGCAGTTGAGACCG | Gpr4 | NM_199058 | no | 1184 | 49 | 1 | Not effective | 1.1611939818 | 18 | No | -0.0010509864 | No annotation in assembly used | | No | No |
| 448 | TGAGTGAAGCAGTTGAGACGC | Gpr4 | NM_199058 | no | 5154 | 177 | 200 | Effective | 1.1611939818 | 157 | No | -0.0010509864 | No annotation in assembly used | | No | No |
| 449 | GCCTTCTTCACATCAGCGC | Sirt4 | NM_201396 | no | 4090.5 | 135 | 0 | Not effective | 0.8441818789 | 313 | No | -0.0045563554 | 0 | 0 | No | No |
| 450 | TTAAATAATCTCTCAGCTCT | Tcf12 | NM_207653 | no | 11215 | 97 | 0 | Not effective | 0.3137724822 | 373 | No | -0.0059293551 | 1 | 1 | No | No |
| 451 | CGAGAAATCCCCACCTTCT | Tcf12 | NM_207653 | no | 5411 | 149 | 0 | Not effective | 1.1043268625 | 157 | No | -0.0078823495 | 1 | 1 | No | No |
| 452 | GGCGTTGGCGCTGCGCTCCTC | Sirt3 | NM_153960 | no | 5455.5 | 262 | 147 | Effective | 1.1453127789 | 74 | No | 0.0613599643 | 1 | 1 | No | No |
| 453 | TAACGAATGCGCGCGCTCGCT | Sirt3 | NM_153960 | no | 3741 | 262 | 0 | Not effective | 0.8513599643 | 145 | No | 0.0613599643 | 0 | 0 | No | No |
| 454 | TTATGCATGCGAGCGTTCT | Sirt3 | NM_153960 | no | 5184 | 231 | 0 | Not effective | 0.5325823267 | 545 | No | 0.0613599643 | 0 | 0 | No | No |
| 455 | GCTCGCGCGCGCGCAGCGAGC | Zscan29 | NM_001278030 | no | 7953 | 412 | 91 | Effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 456 | GCTCGCGCGCGCGCAGCGAGC | Zscan29 | NM_001278030 | no | 5588.5 | 412 | 0 | Not effective | 0.8513599643 | 367 | No | 0.0613599643 | 0 | 0 | No | No |
| 457 (or Zscan4 sgRNA 1) | AATGCATAATCTCCCACCC | Zscan4c, Zscan4d, Zscan4f | NM_001013765 | no | 5588.5 | 264 | 0 | Not effective | 0.2825290161 | 345 | No | -0.0010833983 | 0 | 0 | No | No |
| 458 (or Zscan4 sgRNA 2) | TGGTTCCTTTATAGCCCC | Zscan4c, Zscan4d, Zscan4f | NA_001013765 | no | 4951.5 | 351 | 0 | Not effective | 0.5403154184 | 335 | No | -0.0010833983 | 0 | 0 | No | No |
| 459 (or MERVL sgRNA 1) | AATGACAATCCTCCCACCC | MERVL LTRs | NA | Multi-mapping | 1355 | 372 | 1st assessed in none | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | No |
| 460 (or MERVL sgRNA 3) | GCTCACAACTCAGCTTCAGCCC | MERVL LTRs | NA | Multi-mapping | 5088.5 | 1824 | 1st assessed in none | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | Not assessed in screen | No |
| 461 (or H7 1) | ATGCTTGCAGCTTGCTGCGG | NA | NA | non-targeting | 3572.5 | 286 | 0 | Not effective | 0.3525823267 | 545 | No | 0.0613599643 | 0 | 0 | No | No |
| 462 (or H7 2) | CGCGCCCGTTAGGGAGAGAG | NA | NA | non-targeting | 6184 | 314 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 463 | ATTGTTGAACGCTTACGCG | NA | NA | non-targeting | 6909 | 314 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 464 | GCTTCTACTGCGCAACGTATT | NA | NA | non-targeting | 3653.5 | 274 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 465 | AGCTAGTCTACGTATTCTAT | NA | NA | non-targeting | 5967.5 | 130 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 466 | TAGAATTATACGTCGCCGTG | NA | NA | non-targeting | 2238 | 130 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 467 | ACTTCTATACGTCGCCGTG | NA | NA | non-targeting | 1781 | 130 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 468 | TACTCTATACGTCGCCGTG | NA | NA | non-targeting | 4053.5 | 130 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 469 | CTCGGCCTATTCAGCGATAG | NA | NA | non-targeting | 3486 | 114 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 470 | GCGCTTACCCGCGAAAACCAT | NA | NA | non-targeting | 3357.5 | 219 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 471 | ATAGATTCTCACCCCCGTT | NA | NA | non-targeting | 4053.5 | 219 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 472 | TCTCGCCTATTCAGCGATAG | NA | NA | non-targeting | 3357.5 | 114 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 473 | CTCGGCCTATTCAGCGATAG | NA | NA | non-targeting | 3357.5 | 114 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 474 | GCGCTTACCCGCGAAAACCAT | NA | NA | non-targeting | 3357.5 | 219 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |
| 475 | TCTCAGCTTCGTTAGCGATAGGA | NA | NA | non-targeting | 2278.5 | 130 | 0 | Not effective | 0.8513599643 | 309 | No | 0.0613599643 | 0 | 0 | No | No |

# Appendix D cDNA sequences

Sequences cloned into pIG400 without stop codon and used for cDNA overexpression in the experiments described in sections 5.2.3 and 5.2.4

***Dppa2***:

atgtcatacttcggcctggagactttcaacgagaaccaatctgaggaaaacttggatgaagaaagtgtgattttaacgctggtccccctttaaggagga
ggaggagccaaacacagactacgctacgcaatcaaatgtttcttcttcaaccttagaccacacaccaccagcccgttctctggtgaggcatgctgga
atcaaacacccaaccagaaccataccaagcacctgtcctccgcccagcctgcctcctattagggacgtgtcccggaacactcttcgggagtggtgt
cggtatcataacttgagtacggatggcaagaaagttgaggtctatttgagacttcggagacactcctattctaaacaagaatgttatattcccaatacat
ctcgggaggccagaatgaagcagggtcccaagaaatccaagatagtcttcagaggaatcgggcctccaagcgggtgccaaaggaaaaaggag
gagagtggtgtcctggaaattctaacttcacctaaggagtccacatttgcagcctgggcaaggattgccatgagagcagctcagtcaatgtctaaga
atcgatgtcctcttccgtctaatgtggaggcctttctgccacaagccactggatccagatggtgtgttgtccatggcaggcagctccctgcagataag
aaaggttgggttcgcctgcaatttcttgctggacagacctgggttccagacactccccaaaggatgaattttctcttcctgttaccggcctgtattatccc
agaaccaggagtggaagataatttgctatgtcctgaatgtgttcacagcaacaagaagatcctaagaaactttaaaattagaagccgtgcaaagaaa
aatgccctacccccaaatatgccgccg

***Smarca5:***

atgtcgtccgcggtggagcctccgccgcccccgcctcccgagagcgcgccttccaagccctcggcggcgggcgccggcgggagcagcagcg
gcaacaaaggcggccccgagggcggcgcggcgcccgcggctccgtgtgctgcgggctcgggccccgcggacaccgagatggaggaagtat
ttgatcatggatcacctggaaagcaaaaagaaatccaagaaccagatcctacatatgaagaaaaaatgcaaactgaccgagcaaatagatttgagt
atttattaaagcagacagagctgttcgcacatttcattcagcctgctgctcagaagactccaacctcacccttgaagatgaaaccagggcgcccacg
ggtaaaaaaggatgaaaaacagaatttgctttcagttggagattaccgacaccgtagaacggagcaagaggaggatgaagagctattaacagaaa
gttccaaagcaactaatgtttgtactcgatttgaagattctccatcatatgtaaaatggggtaaactgagagattatcaggtgcgaggattgaattggct
catctctctgtatgagaatggcatcaatgggatccttgcagatgaaatgggtttgggaaagacacttcaaacaatttctcttcttggatacatgaaacac
tatagaaatattcctggtcctcatatggttttggttcccaagtctacattgcacaactggatgagtgaattcaagaaatgggtaccaacacttaggtctgtt
tgcttgataggagataaagagcaaagagctgcatttgtcagggatgttttattaccaggagagtgggatgtgtgtgtcacatcttatgaaatgcttatta
aagagaagtctgtgttcaaaaaattcaactggagatacttagttatagatgaagctcacaggataaaaaatgaaaaatctaagttgtcagaaatagtga
gagaattcaagactacaaatcggctgttgctaacagggacgcctcttcagaacaacctgcatgagctctggtcacttctaaacttttgttaccagatgt
gtttaattcagctgatgactttgattcctggtttgatacaaataattgccttggggatcaaaagctagttgagaggcttcatatggttttgcgtccattcctc
cttcgtcgaattaaagctgatgttgaaaagagtttgcctccaaaaaaagaagtaaaaatctatgtgggtcttagcaaaatgcaaagagaatggtatact
cggatattaatgaaagatatagacatattaaactcagcagggaagatggacaaaatgaggctattgaacattttgatgcagttgaggaaatgctgcaa
tcatccttatctcttcgatggagctgaacctggtccaccgtacacaacagatatgcatctagttaccaacagtggcaagatggtggtgttagacaagct
actccctaaactgaaagaacaaggttcaagagtactaatctttagtcagatgacaagagtattagacattttggaagattattgcatgtggagaaattat
gagtactgcaggttggatggacagacaccccatgatgagagacaagactccatcaatgcatacaatgaaccaaatagcacaaagtttgtatttatgct

aagcacacgtgctggtggtctgggcatcaatcttgcaactgctgatgtagtcattttatatgactcagattggaatccgcaagttgatcttcaggctatg
gaccgtgcccatagaattggacaaacgaagactgtccgagtgtttcgctttataactgataacactgtggaagaaagaatagtggaacgtgctgaga
tgaaacttagactggattcgatagtaattcaacaagggaggcttgtagatcagaacctgaacaaaattgggaaagatgaaatgcttcaaatgattcga
catggtgctacacatgtgtttgcttcaaaagaaagtgaaatcactgatgaggatattgatggtattttagaaagaggtgcaaagaagactgcagaaat
gaacgaaaagctctccaagatgggtgaaagttcacttagaaactttacgatggacacagagtccagtgtttataactttgaaggagaagactatagag
aaaaacaaaagattgcattcacagagtggattgaaccaccaaaaagagaaagaaaagccaactatgctgttgatgcatatttcagggaagctctccg
tgttagtgagcctaaagcacccaaggcacctcgacctccaaaacaaccaaatgttcaggatttccagttctttcctccgcgtttatttgaattactggaa
aaagaaattctgtattacaggaaaactattgggtataaggtacctcggagtcctgatctacctaatgcagcacaggcacagaaagaggaacagctaa
aaattgatgaagcagagcccctcaatgacgaggagttagaggaaaaagagaagcttctgacacaggggtttacaaattggaataagagagatttta
accagtttatcaaggctaatgagaagtggggtcgtgatgatattgaaaatatagccagagaagtagaaggaaaaactccagaagaagtcattgaata
ctcagctgtgttctgggaaagatgcaatgagcttcaggacatagagaagattatggctcagattgaaaggggagaggcaagaattcaaagaagaat
aagtatcaagaaagcactagatacaaagattggacggtacaaagcacccttcatcagctaagaatatcatatggtactaacaaaggaaaaaactat
actgaagaagaggatcgttttctcatttgtatgctgcacaagcttggcttcgacaaggaaaatgtgtatgatgaattacgacaatgcatccgaaactcc
cctcagtttagatttgattggtttctcaagtccagaactgcgatggagctccagaggaggtgtaatacctttaattactttgattgaaagagaaaatatgg
aactagaagaaaaggagaaggcagagaaaaagaaaagaggaccaaaaccttccacacagaaacgtaagatggacggtgcccctgatggccgt
ggaagaaaaaagaagctgaaacta

***Patz1:***

atggagcgggtcaacgacgcttcttgcggtccgtcgggctgctacacctaccaggtgagcagacacagtacggagatgctgcacaacctgaacc
aacaacgcaaaaacggcgggcgcttttgcgacgtgctcctacgagtaggcgacgagagcttcccagcgcaccgcgccgtactggctgcctgcag
cgagtactttgagtctgtgttcagcgcccagttaggcgacggcggagctgcagatggtggtcctgctgatgtgggaggcgcggcggcggctcca
ggcggcggagctggaggcagccgcgaactggagatgcacaccatcagttccaaagtgttcggagacatcctggacttcgcttatacgtcccgaat
cgttgtgcgcctagagagcttccccgagctcatgacggccgccaagttcctgctgatgaggtcggtcatcgagatctgccaggaagtaatcaaaca
gtccaacgtgcagatcctcgtgccccctgcccgggctgatatcatgctctttcgcccacctgggacttctgacttgggcttccctttggacatgaccaa
cggggcagccatggcagccaacagtaacggtattgctggcagtatgcagcccgaggaggaggctgccagggccacaggtgctgctattgcggg
ccaagcttctctgcctgtgttacccggggtggacagattgcccatggtggctggacccctatccccccaactactgacttctccgttccctaatgtggc
atccagtgcacctccactaactagcaagcgaggccggggacgccccaggaaggccaacctgctggactccatgtttgggtctccaggggcttg
agggaagcaggcatccttccatgtggcctgtgcgggaaggtgttcactgacgccaaccggctccggcaacatgaggcccagcacggcgtcaca
agcctccagttgggctatatcgatcttcctcctccaaggctgggtgagaatgggttacccatctccgaggaccccgatggccccagaaaaaggagc
cggaccaggaagcaagtggcttgtgagatctgtggcaagatctttcgtgacgtataccatctcaaccggcataagctttcccactcgggggagaag
ccgtactcgtgcccggtgtgtggtctgcggttcaagagaaaagaccgaatgtcgtaccatgtgaggtcccatgatgggtcagtggggaagtacttg
cgggccgcatacatggcagaccacctgaagaagcacagtgaggggcccagcaacttctgtagcatctgtaaccgagaaggccagaaatgctcac
atcaggatctgattgagagctccgactcctacggtgacctctccgacgccagcgacctgaagacgccagagaaacagagtgccaacggctccttc
tcctgcgacgtggcagtccctaaaaacaaaatggagtctgacggggagaagaagtacccatgccctgaatgtgggagcttcttccgttctaagtcct
acttgaacaaacacatccagaaggtgcatgtccgtgcccttgggggtcccttgggggacctgggccctgcccttggttcacctttctctccccagca
gaacatgtctcttcttgagtcctttggatttcagattgttcagtcagcgtttgcatcatctttagtggatcctgaggtggaccagcagcccatggggcctg
aagggaag

***Carhsp1:***

atgtcatctgagcctcctccaccgccgctgcagccccccacgcatcagacttctgtagggctattggataccctcgcacccgggatcgctctccat
ccccctgcggggcaatgtggtccctagccctctgcctacccgccggacaaggaccttctcagcgacagtgcgggcttcgcagggccccgtctac
aaaggagtctgtaaatgcttctgccggtccaagggccacggcttcatcaccccagctgatggtggtcctgacatcttcctgcacatctctgatgtgga

aggggagtatgttccggtggaaggcgacgaagtcacctataagatgtgctctatcccacccaagaacgagaagctgcaggctgtagaggtggtca
tcacccacctggcaccagggaccaagcatgagacctggtctgggcacgtcatcagcaac

Sequences cloned into pIG400 without stop codon and used for cDNA overexpression in the validation experiments of Appendix J:

***Cbx2:***

atggaggagctgagcagcgtgggcgagcaggtcttcgccgccgagtgcatcctgagcaagcggctccgcaagggcaagctggagtacctggtc
aagtggcgcggctggtcctccaaacacaacagctgggagccagaagagaacattttggacccgaggctgctcctagccttccagaagaaggaac
atgagaaggaggttcagaaccggaagagaggcaagagacccaggggcaggccgaggaaacacacagtcacatcctcctgcagccggcgctc
caagctcaaggaaccagatgcgccatccaaatccaaatccagcagttcgtcctcttcctccacatcttcctcctcttcctcggacgaagaggaagac
gacagcgacctagactccaagagaggcccccggggccgtgaaacccatccagtgcctcagaaaaaagcccagatcctggtagccaagccaga
gctgaaggatcccattagaaagaaacggggacgcaagcctctacccccagaacagaaggcagctcggagacccgtcagcctggccaaggtgct
aaagaccaccaggaaggatctggggacctcagccgccaagctgccccctccactcagcgctccggtggcaggcctggctgccctgaaggccca
caccaaagaggcctgtggtggccccagcactatggcgaccccagagaacctggccagtctgatgaaaggcatggccgggagccccagcagag
gcggcatctggcagagctccatcgtacactacatgaaccgcatgagccagagtcaggttcaggctgccagccgactggcactcaaggcccaggc
caccaacaagtgcggtctcgggctagacctgaaagtgaggacgcagaaggggggtgagctaggggggagccccgcaggaggcaaggtccc
gaaggcccccggtggcggagctgcagagcagcagagagggaaccattcggggagcccaggtgctcagctggcacccactcaggagttgagc
cttcaggtccttgacttgcaaagcgtcaagaacggtgtgcctggtgtgggcctgcttgctcgccatgccccagccaaggctattcctgctaccaacc
cagccacagggaaaggtcctgggagcggccccacaggagcaaacatgaccaacgctcccacagacaacaacaaaggggaaaagctgacttg
caaggcaacggctctgcctgccccttccgtcaagcgggacaccgttaaaagcgtcgctgcctccggcgggcaggagggccacacagccccgg
gagaaggccgaaagccacctgcgctgtctgagctgagcacgggagaggagaatagtagctctgactcggaccctgactcgacctcgcttcccag
tgctgcgcagaacctatctgtagctatccagaccagccaggactggaaacctacccgcagtctcatcgaacacgtctttgtcacggatgtcacagcc
aacctcatcactgtcaccgtgaaggagtcgcccaccagcgtgggcttcttcaacttgaggcattat

***Dppa3:***

atggaggaaccatcagagaaagtcgacccaatgaaggaccctgaaactcctcagaagaaagatgaagaggacgctttggatgatacagacgtcc
tacaaccagaaaacactagtaaaggtcatgaaaaagctaaccctaaaccccggtgtcaagcggtccgcacgccggcgcagtctacggaaccgcatt
gcagccgtacctgtggagaacaagagtgaaaaaatccggagggaagttcaaagcgcctttcccaagagaagggtccgcactttgttgtcggtgct
gaaagaccctatagcaaagatgagaagacttgttcggattgagcagagacaaaaaaggctcgaaggaaatgagtttgaacgggacagtgagcca
ttcagatgtctctgcactttctgccattatcaaagatgggatccctctgagaatgcgaaaatcgggaagaat

***Dppa4:***

atggagactgctggagacaagaagtggagcgcagaagagccgaaggaagaagtggaattgcagatgtctagtcaaccaagcacggctcctgca
aaggctaaagcaacggggaaaaaacaaaagaagtcggagacagataatggttgtaaaccaaaggagggaaaaccacaagacactgagacgcc
aggacagactcgtaggaagctggagttgaggaacgtccctgactcagccaaagacgccaggttgaagacagctcacaaaaaaatgaagactgaa

ccgggggaggagtctgaggtgacagttcctctggaaatggtccctgtgccagaggagcagatacctgccctcattgaccctcctatgctctatgag
gaagtcagcaccaccgtagtgactacacctgccactgaggccgtgttagcatcttgggccagaattgcatccaatgctaagaagtacgaggcagtg
ccagccgatgcttcgtcctcatcagaagtcaaaggggaaatgtggtgtgtggttcatgggaccagccttcctgggaactcgcgtggttgggttcggc
tgcagttccacgctggacaagcctgggtacccgataagaaaggaaaagccattgccctcttcctgcttccggcctgcacatttccacccccacacct
ggaggacaacatgctgtgccccaagtgtgttcataagaacaagatcttgactaagagcctcgaagga

***E2f5:***

atggcggcggcggagcccacgagctctgctcagcccacgccgcaggcccaggctcagccgccgccgcatggggcgccatcctcgcagccgt
ccgcggcgctcgcgggggggcagcagccggcacgagaagagcctgggcttgcttaccaccaaattcgtgtcgttgctgcaggaggcgcaggac
ggcgtcctggatctcaaagcggctgcagataccttggctgtgaggcaaaagcgaagaatttatgatatcaccaatgtcttagagggaattgatctaat
tgaaaaaaaatcaaagaacagtatccagtggaagggtgtaggtgctggctgtaatactaaagaagttatcgatagattaaggtgtcttaaagctgaaa
ttgaagatctcgaattgaaggaaagagaacttgaccagcagaagttgtggctacagcaaagcatcaaaaatgtgatggaagactccattaataacag
attttcttatgtaactcacgaagacatctgcaattgctttcatggtgatacactgttggccattcaggcaccttctggtacacagctggaagtacctattcc
agaaatgggacagaatggacaaaagaaataccagataaatctgaagagtcactcagggcctatccatgtgctacttataaataaagagtccagttca
tctaagccagtggtttttcctgttcccccacctgatgacctcacacagccttcctcccagtcctcaacttcagtgactccacagaaatccaccatggctg
ctcaaaacctgcctgagcagcatgtttccgaaagaagccagactttccagcagacaccagctgcagaagtatcttcaggatctattagtggagacat
cattgatgaactgatgtcttctgatgtgtttcctcttttacggctttctcctaccccagcagatgactacaactttaatttagatgataatgaaggagtttgtg
atctgtttgatgttcagatactaaattat

***Ep400:***

atgcaccatggcagtggtcctcagaatgtccaacatcagctgcagaggtctaggtccttcactgggagtgaagaggagcagccagcccatcccaa
cttacccccatctcctgcagcgcctttcgctccatcagccagtccatctgcaccccaatcccctgggtaccagatacagcagctgatgagcagaagt
cctgtggccgggcagaatgtgaacatcacactgcagaatgttggcccagttgtgggaggaaaccagcagatcacactggccccccctgccactgcc
caaccccacttctccaggttttcagtttggtgcccagcaaaggcgctttgagcatggctctccatcatacattcaagttacttctccgatgtcgcagcag
gtccagactcaaagccccacacagcctagccctgggccagggcaaaccctgcagaatgtgcgtgcaggtgctccaggtcctgggctgggcatat
gcagcaacagccccactggaggctttgtggatgccagtgtgcttgtgaggcagatcagcttgagtccatcaagtggtggtcactttgtatttcaggag
gcaccaggcctgactcagatggcccagggagcccaggttcaactccagcattcaggagctcccatcactgtccgagaacggagactctcccaac
ctcatgcacagtcgggaggcaccatccaccatctagggcctcagagccctgcagctgcaggaggcactggacttcaacctctggccagcccaaa
ccacatcaccacggccagcctgccaccccagatcagcagcattattcagggccagttgatacagcaacagcagcaggtgcttcagggacagcca
atgaaccgatctctgggatttgagaggacacctggtgtgctgcttcctggcgtgggagggccttcagcatttgggatgacatccccaccaccaccca
ccagcccttccagaaccaccatgcctccaggcctttccagtgtgcctcttacatctatgggaagctcaggaatgaagaaggttcccaagaagttgga
ggagatcccccccagcctcccaagagatggcacagatgaggaagcagtgtctggactaccactacaaggaaatggaggcactcaaggaagtcttc
aaggagtacctgattgaactgttcttcttgcaacatctacagggaaacatgatggatttcttggcttttaaaaagaagcactatgcccctttacaagcgta
tcttaggcagaatgatttggacattgaagaggaggaggaggaagaggaagaggaggaaggaaaatctgaagttatcaacgatgagcaccaatcc
ctgactgggagcctagtggtggggcctggcagtgccacagaagccgaccccttaagaggcaacaggtgatgccacccacagagcagtctaag
agacctcgtcttgaagtgggtcatccaggggtagttttccagcatccaggggtgaatgcaggagtccctctgcagcagttaatgccaacagtccaag
gagggatgcccccaactcctcaggccacacagctcactggacagaagcagagtcagcagcagtacgacccttccacagggcctcctgtgcaga
acgccgccagcttgcataccccaccaccacagcttcctgcgaggctgcctccggccagtgtcccagccacagccctcccttccaccttacagttttc
acagcagtcacagatggtggaggcttcaactcagctccaaatccctgtgaagactcaacagctaaatgcccccatccctgcccccctgcccagcca
gctccctgctccttcttcacagcccgcccagccggccctacatgttccgatgcccgggaaggcgcagatgcagacatctcagctttcctcccagac
ccaggtactctgtggacattggctgttttac

***Ezh2:***

atgggccagactgggaagaaatctgagaagggaccggtttgttggcggaagcgtgtaaaatcagagtacatgagactgagacagctcaagaggt
tcagaagagctgatgaagtaaagactatgtttagttccaatcgtcagaaaatttttggaaagaactgaaaccttaaaccaagagtggaagcagcggag
gatacagcctgtgcacatcatgacttctgtgagctcattgcgcgggactagggagtgttcagtcaccagtgacttggatttttccagcacaagtcatcc
cgttaaagaccctgaatgcagtcgcctcggtgcctataatgtactcttggtcgcccttacaacagaattttatggtggaagacgaaactgttttacataa
cattccttatatgggggatgaagttctggatcaggatggcactttcattgaagaactaataaaaaaattatgatggaaaagtgcatggtgacagagaatg
tggatttataaatgatgaaatttttgtggagttggtaaatgctcttggtcaatataatgatgatgatgatgacgatgatggagatgatccagatgaaagag
aagaaaaacagaaagatctagaggataatcgagatgataaagaaacttgcccacctcggaaatttcctgctgataaaatatttgaagccatttcctca
atgtttccagataagggcaccgcagaagaactgaaagaaaaatataaagaactcacggagcagcagctcccaggtgctctgcctcctgaatgtact
ccaaacatcgatggaccaaatgccaaatctgttcagagggagcaaagcttgcattcatttcatacgctcttctgtcgacgatgttttaagtatgactgctt
cctacatcccttccatgcaacacccaacacatataagaggaagaacacagaaacagctttggacaacaagccttgtggaccacagtgttaccagca
tctggagggagctaaggagtttgctgctgctcttactgctgagcgtataaagacaccacctaaacgcccaggggccgcagaagaggaagacttc
cgaataacagtagcagacccagcacccccaccatcagtgtgctggagtcaaaggatacagacagtgacagagaagcagggactgaaactgggg
gagagaacaatgataaagaagaagaagagaaaaaagatgagacgtccagctcctctgaagcaaattctcggtgtcaaacaccaataaagatgaa
gccaaatattgaacctcctgagaatgtggagtggagtggtgctgaagcctccatgtttagagtcctcattggtacttactacgataacttttgtgccattg
ctaggctaattgggaccaaaacatgtagacaggtgtatgagtttagagtcaaggagtccagtatcatagcacctgttcccactgaggatgtagacact
cctccaagaaagaagaaaaggaaacatcggttgtgggctgcacactgcagaaagatacaactgaaaaaggacggctcctctaaccatgtttacaa
ctatcaaccctgtgaccatccacggcagccttgtgacagttcgtgcccttgtgtgatagcacaaaatttttgtgaaaagttttgtcaatgtagttcagagt
gtcaaaaccgctttcctggatgtcggtgcaaagcacaatgcaacaccaaacagtgtccatgctacctggctgtccgagagtgtgaccctgacctctg
tctcacgtgtggagctgctgaccattgggacagtaaaaatgtatcctgtaagaactgtagcattcagcggggctctaaaaagcacttactgctggcac
cgtctgatgtggcaggctggggcatctttatcaaagatcctgtacagaaaaatgaattcatctcagaatactgtggggagattatttctcaggatgaag
cagacagaagaggaaaagtgtatgacaaatacatgtgcagctttctgttcaacttgaacaatgattttgtggtggatgcaacccgaaagggcaacaa
aattcgttttgctaatcattcagtaaatccaaactgctatgcaaaagttatgatggttaatggtgaccacaggataggcatctttgctaagagggctatcc
agactggtgaagagttgttttttgattacagatacagccaggctgatgccctgaagtatgtgggcatcgaacgagaaatggaaatccct

***Hmgb2:***

atgggcaagggtgaccccaacaagccgcggggcaaaatgtcctcgtacgccttcttcgtgcagacctgccgcgaggagcacaagaagaagcat
cccgactcgtcggtgaacttcgccgagttctccaagaaatgctccgagagatggaagaccatgtctgcaaaggaaaagtccaagtttgaagatttgg
ccaagagcgacaaagctcgttatgacagggagatgaagaactatgttcctcccaaaggggataagaaaggaaagaaaaaagaccccaatgctcc
gaagagaccaccgtctgccttcttcctgttttgctctgaaaatcgcccaaagatcaaaattgaacacccaggcctgtctattggagatactgcgaaga
aactgggtgagatgtggtctgagcaatctgccaaagataaacaaccgtatgagcagaaagcagctaaactaaaggagaagtatgaaaaggatatt
gctgcataccgtgccaagggcaaaagtgaagcaggaaagaagggtcctggtaggccaacaggctcaaagaagaagaacgaaccagaagatga
ggaggaggaagaagaggaggaagaggaggaagatgacgaggaagaagaggaggatgaagaa

***Lin28a:***

atgggctcggtgtccaaccagcagtttgcaggtggctgcgccaaggcagcggagaaggcgccagaggaggcgccgcctgacgcggcccgag
cggcagacgagccgcagctgctgcacggggccggcatctgtaagtggttcaacgtgcgcatggggttcggcttcctgtctatgaccgcccgcgct
ggggtcgcgctcgaccccccggtggacgtctttgtgcaccagagcaagctgcacatggaagggttccgaagcctcaaggagggtgaggcggtg
gagttcacctttaagaagtctgccaagggtctggaatccatccgtgtcactggccctggtggtgtgttctgtattgggagtgagcggcggccaaaag
ggaagaacatgcagaagcgaagatccaaaggagacaggtgctacaactgcggtgggctagaccatcatgccaaggaatgcaagctgccacccc
agcccaagaagtgccacttttgccaaagcatcaaccatatggtggcctcgtgtccactgaaggcccagcagggccccagttctcagggaaagcct
gcctacttccgggaggatgaggaagagatccacagccctgccctgctcccagaagcccagaat

***Ncoa3:***

atgagtggactaggcgaaagctctttggatccgctggccgctgagtctcggaaacgcaaactgccctgtgatgccccaggacaggggcttgtctac
agtggtgagaagtggcgacgggagcaggagagcaagtacatagaggagctggcagagctcatctctgcaaatctcagcgacatcgacaacttca
atgtcaagccagataaatgtgccatcctaaaggagacagtgagacagatacggcaaataaaagaacaaggaaaaactatttccagtgatgatgatg
ttcaaaaagctgatgtgtcttctacagggcagggagtcattgataaagactctttaggaccgcttttactacaggcactggatggtttcctgtttgtggtg
aatcgagatggaaacattgtattcgtgtcagaaaatgtcacacagtatctgcagtacaagcaggaggacctggttaacacaagtgtctacagcatctt
acatgagcaagaccggaaggattttcttaaacacttaccaaaatccacagttaatggagtttcttggactaatgagaaccagagacaaaaaagccata
catttaattgtcgtatgttgatgaaaacacacgacattttggaagacgtgaatgccagtcccgaaacacgccagagatatgaaacaatgcagtgctttg
ccctgtctcagcctcgcgctatgctggaagaaggagaagacttgcagtgctgtatgatctgcgtggctcgccgcgtgactgcgccattcccatccag
tcctgagagctttattaccagacatgacctttccggaaaggttgtcaatatagatacaaactcacttagatcttccatgaggcctggctttgaagacata
atccgaagatgtatccagaggttctttagtctgaatgatgggcagtcatggtcccagaagcgtcactatcaagaagcttatgttcatggccacgcaga
gacccccgtgtatcgtttctccttggctgatggaactattgtgagtgcgcagacaaaaagcaaactcttccgcaatcctgtaacgaatgatcgtcacg
gcttcatctcgacccactttcttcagagagaacagaatggatacagaccaaacccaaatcccgcaggacaaggcatccgacctcctgcagcaggg
tgtggcgtgagcatgtctccaaatcagaatgtacagatgatgggcagccggacctatggcgtgccagaccccagcaacacagggcagatgggtg
gagctaggtacggggcttctagtagcgtagcctcactgacgccaggacaaagcctacagtcgccatcttcctatcagaacagcagctatgggctca
gcatgagcagtccccccccacggcagtcctggtcttggtcccaaccagcagaacatcatgatttcccctcggaatcgtggcagcccaaagatggcct
cccaccagttctctcctgctgcaggtgcacactcacccatgggaccttctggcaacacagggagccacagcttttctagcagctccctcagtgccttg
caagccatcagtgaaggcgtggggacctctcttttatctactctgtcctcaccaggccccaaactggataattctcccaatatgaatataagccagcca
agtaaagtgagtggtcaggactctaagagccccctaggcttatactgtgaacagaatccagtggagagttcagtgtgtcagtcaaacagcagagat
cacccaagtgaaaaagaaagcaaggagagcagtggggaggtgtcagagacgcccaggggacctctggaaagcaaaggccacaagaaactgc
tgcagttactcacgtgctcctccgacgaccgaggccattcctccttgaccaactctcccctggatccaaactgcaaagactcttccgttagtgtcacca
gcccctctggagtgtcctcctcaacatcagggacagtgtcttccacctccaatgtgcatgggtctctgttgcaagagaaacaccggattttgcacaag
ttgctgcagaatggcaactccccagcggaggtcgccaagatcactgcagaggccactgggaaggacacgagcagcactgcttcctgtggagag
gggacaaccaggcaggagcagctgagtcctaagaagaaggagaataatgctctgcttagatacctgctggacagggatgaccccagtgatgtgc
ttgccaaagagctgcagccccaggccgacagtggggacagtaaactgagtcagtgcagctgctccaccaatcccagctctggccaagagaaaga
ccccaaaattaagaccgagacgaacgaggaggtatcgggagacctggataatctagatgccattcttggagatttgaccagttctgacttctacaac
aatcctacaaatggcggtcacccaggggccaaacagcagatgtttgcaggaccgagttctctgggtttgcgaagtccacagcctgtgcagtctgttc
gtcctccatataaccgagcggtgtctctggatagccctgtgtctgttggctcaggtccgccagtgaagaatgtcagtgctttccctgggttaccaaaac
agcccatactggctgggaatccaagaatgatggatagtcaggagaattacggtgccaacatgggcccaaacagaaatgttcctgtgaatccgactt
cctcccccggagactggggcttagctaactcaagggccagcagaatggagcctctggcatcaagtcccctgggaagaactggagccgattacag
tgccacttttacccagacctgccatgggggctctgtgcctaccttgccacttcgttctaatcgactgccaggtgcaagaccatcgttgcagcaacagc
agcagcaacagcagcaacagcaacaacaacagcagcaacagcagcagcaacagcagcagcagcaacagcagcagatgcttcaaatgagaac
tggtgagattcccatgggaatgggagtcaatccctatagcccagcagtgccgtctaaccaaccaggttcctggccagagggcatgctctctatgga
acaaggtcctcacgggtctcaaaataggcctcttcttagaaactctctggatgatctgcttgggccaccttctaacgcagagggccagagtgacgag
agagctctgctggaccagctgcacacactcctgagcaacacagatgccactggtctggaggagatcgacagggccttgggaattcctgagctcgt
gaatcagggacaagctttggagtccaaacaggatgttttccaaggccaagaagcagcagtaatgatggatcagaaggctgcactatatggacaga
catacccagctcagggtcctccccttcaaggaggctttaaccttcagggacagtcaccatcgtttaactctatgatgggtcagattagccagcaaggc
agctttcctctgcaaggcatgcatcctagagccggcctcgtgagaccaaggaccaacaccccgaagcagctgagaatgcagcttcagcagaggc
tacagggccagcagtttttaaatcagagccggcaggcacttgaaatgaaaatggagaaccctgctggcactgctgtgatgaggcccatgatgcccc
agcaggctttctttaatgcccaaatggctgcccagcagaaacgagagctgatgagccatcacctgcagcagcagaggatggcgatgatgatgtca
caaccacagcctcaggccttcagcccacctcccaacgtcaccgcctcccccagcatggacggggttttggcaggttcagcaatgccgcaagcccc
tccacaacagtttccatatccagcaaattacggaatgggacaaccaccagagccagcctttggtcgaggctcgagtcctcccagtgcaatgatgtca
tcaagaatggggccttcccagaatgccatggtgcagcatcctcagcccacacccatgtatcagccttcagatatgaaggggtggccgtcagggaa
cctggccaggaatggctccttcccccagcagcagtttgctccccaggggaaccctgcagcctacaacatggtgcatatgaacagcagcggtgggc
acttgggacagatggccatgacccccatgcccatgtctggcatgcccatgggccccgatcagaaatactgc

***Plac8:***

atggctcaggcaccaacagttatcgtgactcaacctggattcgttcgtgctccccaaaattccaactggcagaccagcctgtgtgattgcttcagtgac
tgcggagtctgcctctgtgggacctttgtttcacttgtcttggatgtcaagtggcagctgacatgaatgagtgttgtctgtgtggaacaacggtggcca
tgaggactctctaccgaacccgatacggcattcctggatctatttgtgatgactacatggtcacactcttctgtcctgtttgctctgtgtgccaactcaag
agagacattaacaggaggagagccatgaacgctttc

***Pou5f1:***

atggctggacacctggcttcagacttcgccttctcacccccaccaggtgggggtgatgggtcagcagggctggagccgggctgggtggatcctcg
aacctggctaagcttccaagggcctccaggtgggcctggaatcggaccaggctcagaggtattggggatctccccatgtccgcccgcatacgagt
tctgcggagggatggcatactgtggacctcaggttggactgggcctagtcccccaagttggcgtggagactttgcagcctgagggccaggcagg
agcacgagtggaaagcaactcagagggaacctcctctgagccctgtgccgaccgccccaatgccgtgaagttggagaaggtggaaccaactcc
cgaggagtcccaggacatgaaagccctgcagaaggagctagaacagtttgccaagctgctgaagcagaagaggatcaccttggggtacaccca
ggccgacgtggggctcaccctgggcgttctctttggaaaggtgttcagccagaccaccatctgtcgcttcgaggccttgcagctcagccttaagaac
atgtgtaagctgcggcccctgctggagaagtgggtggaggaagccgacaacaatgagaaccttcaggagatatgcaaatcggagaccctggtgc
aggcccggaagagaaagcgaactagcattgagaaccgtgtgaggtggagtctggagaccatgtttctgaagtgcccgaagccctccctacagca
gatcactcacatcgccaatcagcttgggctagagaaggatgtggttcgagtatggttctgtaaccggcgccagaagggcaaaagatcaagtattga
gtattcccaacgagaagagtatgaggctacagggacacctttcccaggggggctgtatcctttcctctgcccccaggtccccactttggcacccca
ggctatggaagcccccacttcaccacactctactcagtccctttcctgagggcgaggcctttccctctgttcccgtcactgctctgggctctcccatgc
attcaaac

***Preb:***

atgggtcggcgccggggtgtggagctgtaccgggccccgttcccgttgtacgcgcttcggatagaccccaagactgggctgctcatcgctgcgg
gcggaggaggagctgccaagaccggcataaagaatggcgtgcattttctgcagctagagctgatcaacgggtgcctgagcgcttccttgctgcac
tctcatgacacggagacacgggccaccatgaatttggcgcttgctggtgacattcttgctgccggacaggatgcccagtgtcagcttcttcgtttcca
ggtccatcaacagaagggcagtaaagcggagaagtcaggttccaaggagcagggacctcgacagagaaaggggctcctccagcagagaag
aaatcgggagcacaagttcaccccgaaggggttgaactcaaagtaaagaatttggaggcagtacagacagacttcagcaatgaaccgctgcaga
aagttgtgtgcttcaaccatgataacaccctgcttgccactggaggaactgatggtcatgttcgtgtctggaaggtacctagcctggagaaagttctg
gagttcaaagcccatgaaggggagattggagatttgactttgggtcctgatggcaagctggttacagtgggctgggactttaaggcctctgtgtggc
agaaggatcaactggtgacacagctacagtggcaagagaatggaccagcctcttctaatacaccataccgctaccaggcctgcaggtttgggcag
gttccagatcagctcggtgggctgcgactcttcacagtgcagataccccacaagcgcctacgtcagcccccacccctgctacctcacagcctgggac
agttccaccttcttgcctcttcggaccagatcctgtggccatgaggtcatttcctgcctcagtgtcagtgattcgggtaccttcctaggcctaggcacgg
tcactggctctgtcgccatctacatagctttctctctccagcgcctgtattatgtgaaggaggcccatggcattgtggtaacagatgtgaccttttctacct
gagaagggttgtggtccaaagctccttgggccccatgaaactgctcttttctctgtggctgtggatagtcgttgccagttgcacctgctgccttcacgg
cggagtgttcctgtatggcttctgctcctgctgtgtgtcggccttattattgtgaccatcctgctgctccagactgccttcccgggatttctc

ctgcgtcaccattcctcgctccctggatggcaggttgcaggtgtcccaccggaagggactacctcatgtcatttattgccgtgtgtggcgctggcccg
acctccagagccaccatgaactgaagcctctggaatgctgtgagttcccatttggttccaagcagaaggaggtctgcatcaacccctaccactataa
gcgagtggagagcccggttctcccgccggtgctggttccgaggcacagcgagtacaatcctcagcacagccttctggctcagttccgcaacctgg
gacaaaatgagcctcacatgccactgaacgccacgttcccagactctttccagcagcccaacagccacccgttcccccactcccccaacagcagc
taccccaactctcctggcggcagcagcagcacctaccctcactccccaaccagctcagacccgggcagccctttttcagatgccagctgacacacc
cccacctgcttacctgcctcctgaagaccccatggcccaggatggctctcagcccatggacacgaacatgatggcgcctccactgcccgctgaaat
cagcagaggagatgttcaggcagttgcttacgaggaaccaaaacactggtgctctattgtgtactatgagctcaacaaccgtgtgggtgaagcgttc
cacgcctcgtccaccagcgtgctggtggatggtttcacagatccgtccaacaataagaaccgcttctgccttggcttgctctccaacgttaaccggaa
ttccactattgaaaacaccaggcgacatattgggaaaggagtccacctttattacgttggaggagaggtgtatgcggaatgcctcagtgacagcagc
atcttcgtgcagagccggaactgcaactaccaccacggctttcaccccaccaccgtctgcaagatccccagcgggtgcagcttgaaaatcttcaac
aaccaagagtttgctcagctactggcgcagtctgtgaaccacgggttcgagaccgtgtatgaactcaccaaaatgtgcactattcggatgagcttcgt
gaagggttggggagccgaataccaccggcaggatgttaccagcacccccctgctggattgagatccatctgcatggccctctccagtggctggata
aggttctgacccagatgggctcaccccacaatcctatttcatccgtgtct

***Smarca2:***

atgaatcctggttacgaagttgcacccagatctgacagtgaagagagtgaatcggactacgaggaggaggatgaagaagaagagtccagtaggc
aggaaaccgaggagaagatactgctggatcccaacagtgaagaagtttccgaaaaggatgccaagcagatcattgagactgcgaagcaggacgt
ggacgacgaatacagcatgcagtacagtgccagaggctctcagtcctactacacggtggctcacgctatctctgagagggtggagaagcagtctg
ccctcctcattaacggcaccctaaagcattaccagctccagggcctggaatggatggtttccctgtataataacaatctgaacggaatcttagctgatg
aaatggggctaggcaagaccatccagaccattgcactcatcacgtatctgatggagcacaaaaggctcaatggtccctacctcatcatcgtccccct
ctcgactctgtctaactggacatatgaatttgacaaatgggctccttctgtggtgaaaatttcttacaagggtacccctgccatgcgacgctccctcgtt
ccccagctacggagtggcaaattcaatgtcctcctgactacttacgagtacattataaaagacaagcacattcttgcaaagattcggtggaagtacat
gatcgtggacgaaggccaccggatgaagaatcaccactgcaagctaacccaggtcctgaacacacactatgtggcccccaggcggatccttctga
ctgggaccccactgcagaataagcttccggaactctgggccctcctcaacttcctcctccctacaatcttcaagagttgcagcacatttgagcagtgg
tttaatgctccatttgccatgaccggtgaaagggtggacctgaacgaagaagaaacgattttgatcatcaggcgtctacacaaggtgctgagaccctt
tttactgaggaggctgaagaaagaggttgagtctcagcttccggaaaaggttgagtatgtgatcaagtgtgacatgtcagctctgcagaagattctgt
accgtcacatgcaagccaaggggatcctcctcacggacgggtctgagaaagataagaaggggaaaggaggtgccaagacacttatgaacacca
tcatgcagctgagaaaaatatgcaaccacccatatatgtttcagcacattgaggaatcctttgctgaacacctgggctattcgaatggggtcatcaatg
gggctgagctgtatcgggcctcgggaaagtttgagctgcttgatcgtattctgcccaaattgagagcgactaaccaccgcgtgctgcttttctgccag
atgacgtcactcatgaccattatggaggattactttgcttttcggaacttcctgtacctgcgccttgacggcaccaccaagtctgaagatcgtgctgcttt
gctaaagaaattcaatgaacctgggtcccagtatttcattttcttgctgagcacaagagcaggggcctgggcttaaatcttcaggcggcagacacg
gtggtcatatttgacagcgactggaatcctcaccaggatctgcaggcccaagaccgagctcaccgcattggccaacaaaacgaggtccgggtgct
gaggctttgcaccgtcaacagtgtggaggaaaagattctcgcggctgccaagtacaagctgaacgtggatcagaaggttatccaagcaggcatgtt
tgaccagaagtcatccagccacgagcggagggccttcctgcaggccattctggagcacgaggaggagaatgaggaagaagatgaggtaccag
acgacgagaccctgaaccagatgattgctcgccgggaggaagaatttgatctttttatgcgcatggacatggaccggcggagggaggatgcccg
gaacccgaagcgcaaaccccgcttgatggaggaagatgagctgccctcctggattatcaaggatgacgccgaagtggaaaggctcacctgtgaa
gaagaggaggagaagatatttgggaggggctctcgccagcgccgggatgtggactacagtgatgccctcaccgagaagcaatggctcagggcc
atcgaagacggcaatttggaagaaatggaagaggaggtacggcttaagaagagaaaaagacgaagaaatgtggataaagaccccgtgaaggaa
gatgtggaaaaagcgaagaaaagaagaggccgccctccggctgagaagttgtcaccaaatcccccaaaactaacgaagcagatgaacgccatc
attgatactgtgataaactacaaagacagttcagggcgacagctcagtgaagtcttcattcagttaccttccaggaaagacttaccagaatactatgaa
ttaattaggaagccagtggatttcaaaaagataaaggagcgaatccgtaatcataagtatcggagcctgggagacctggagaaagacgtcatgcttc
tctgtcacaacgcacagacattcaacttggaaggatcccagatctacgaagactccattgtcctacagtcagtgtttaagagtgctcggcagaaaatt
gccaaagaagaagagagtgaggaagaaagcaatgaagaagaggaagaagatgatgaagaggagtcggagtcagaggcgaaatctgtgaagg
tgaaaatcaagctgaataaaaaggaagagaaaggccgggacacagggaagggcaagaagcggccaaaccgaggcaaagccaaacccgtcgt
gagcgattttgacagtgacgaggaacaggaagagaacgaacagtcagaagcaagtggaactgataacgag

# Appendix E PCA loading values

The following table contains loading values for PC1 and PC2 for 965 highly-variable genes in the CRISPRa scRNA-seq screen dataset.

| Gene name | PC1 loading value | PC2 loading value |
|---|---|---|
| 0610040J01Rik | 0.015094567 | 0.052495927 |
| 1190005I06Rik | -0.002397842 | -0.019747118 |
| 1500009L16Rik | 0.015628465 | 0.007712967 |
| 1700007K13Rik | 0.002781191 | 0.02246037 |
| 1700029P11Rik | -0.02058883 | 0.023758499 |
| 1700052K11Rik | 0.015544041 | 0.037400838 |
| 1700097N02Rik | -0.037985995 | -0.007069206 |
| 1810032O08Rik | -0.002645922 | -0.0047161 |
| 2200002D01Rik | 0.06666651 | -0.010332194 |
| 2310043M15Rik | -0.00924505 | -0.001522274 |
| 2810417H13Rik | 0.015659865 | -0.049252756 |
| 3830417A13Rik | 0.030609732 | 0.002949171 |
| 4930548H24Rik | -0.006180642 | 0.003333947 |
| 4933402E13Rik | 0.011801084 | 0.007870698 |
| 5430402O13Rik | 0.01647246 | 0.09500527 |
| 5430416N02Rik | 0.002684761 | -0.003469797 |
| 6330410L21Rik | -0.013145719 | 0.021873567 |
| 9530053A07Rik | 0.009405954 | 0.011205982 |
| 9530059O14Rik | 0.017245522 | 0.09756264 |
| A330069E16Rik | -0.000406597 | -0.002033843 |
| A930014E10Rik | 0.007221519 | 0.06124033 |
| AA467197 | 0.00573293 | 0.012449559 |
| Aard | -0.048047964 | 0.041346278 |
| Aass | -0.00342597 | -0.010224617 |
| Abracl | 0.042533085 | -0.007854142 |
| Acaa2 | -0.014266563 | 0.012860991 |
| Acot1 | 0.017250547 | 0.018125627 |

| | | |
|---|---|---|
| Acta1 | 0.029686727 | -0.003736086 |
| Acta2 | 0.03376475 | -0.007041237 |
| Actb | 0.031927373 | -0.028425362 |
| Actg1 | 0.06180474 | -0.039226115 |
| Acy3 | 0.009351596 | 0.054866035 |
| Adamts9 | 0.00251939 | 0.002826694 |
| Adcy2 | 0.008621854 | 0.07617312 |
| Adgrl2 | 0.049820874 | -0.01919885 |
| Adgrl3 | 0.007103615 | 0.037204918 |
| Adprh | -0.047208536 | -0.003131277 |
| Adrb2 | 0.011310952 | 0.0088274 |
| Aebp2 | -0.018760156 | 0.007505677 |
| Aes | -0.064773194 | 0.010382351 |
| AF067061 | 0.013069687 | 0.0813769 |
| Agl | -0.019933186 | 0.006543614 |
| Agpat4 | 0.0435728 | -0.0027023 |
| Agpat9 | 0.025616627 | 0.009243218 |
| Ahnak | 0.045911707 | -0.001916705 |
| Ahnak2 | 0.035194464 | 0.002531802 |
| Aire | -0.040366407 | -0.007281113 |
| Ak4 | -0.012940768 | 0.025580687 |
| Akap12 | 0.018310642 | -0.016972357 |
| Akr1b3 | -0.016898645 | -0.024077412 |
| Aldoa | 0.010901698 | -0.019452162 |
| Amn | 0.008800886 | 0.006152645 |
| Amotl1 | 0.04937692 | -0.00564561 |
| Amotl2 | 0.04054745 | -0.000305379 |
| Amt | -0.004748545 | -0.00376131 |
| Anp32a | -0.022215739 | -0.021950474 |
| Anxa1 | 0.046789225 | 0.002154642 |
| Anxa2 | 0.06540843 | 0.0133765 |
| Anxa3 | 0.109695844 | -0.00502606 |
| Anxa5 | 0.033014603 | 0.007718445 |
| Anxa8 | 0.006824127 | 0.008281822 |
| Ap3b2 | 0.013319263 | 0.005736365 |
| Apela | 0.051585913 | -0.032512095 |
| Apoa2 | 0.003223777 | 0.023951797 |
| Apobec2 | 0.014635964 | 0.010246145 |

| | | |
|---|---|---|
| Apoe | 0.001125963 | 0.014355245 |
| Aqp3 | -0.007551776 | 0.032101583 |
| Arg2 | 0.025373207 | 0.13080893 |
| Arhgef26 | 0.021730153 | 0.096255705 |
| Arl14epl | -0.03452046 | 0.017821318 |
| Arl4a | -0.008026663 | -0.010391519 |
| Arl4c | 0.08486886 | -0.024569554 |
| Arl4d | -0.004283476 | 0.022680787 |
| Arl6ip1 | -0.007842236 | 0.026358351 |
| Arpc1b | 0.020260893 | -0.00683571 |
| Ascl2 | 0.01904236 | 0.007459578 |
| Asns | 0.005183192 | 0.028264124 |
| Ass1 | -0.007555474 | 0.029019162 |
| Atf4 | 0.003308763 | -0.004568338 |
| Atg2a | -0.011278303 | 0.005181089 |
| Atp5b | -0.001511414 | 0.027984366 |
| Atp6v0b | 0.023989808 | 0.005440626 |
| Atp6v0d1 | 0.007337852 | 0.020088376 |
| Atrn | 0.001722835 | 0.01854388 |
| Atrx | -0.001318779 | -0.007089897 |
| Avpi1 | 0.023845786 | 0.058848053 |
| AY036118 | -0.001742739 | 0.00158306 |
| B020031M17Rik | 0.014280799 | 0.09668234 |
| B2m | 0.06365098 | -0.004644726 |
| B930036N10Rik | 0.003977635 | 0.002061978 |
| Bambi | 0.016732853 | 0.016353419 |
| Baz1a | 0.04786045 | -0.001368109 |
| Bbc3 | 0.004886779 | -0.005869682 |
| Bbs2 | -0.0401821 | 0.000656059 |
| Bbx | 0.056650326 | -0.006721663 |
| BC028528 | -0.01344899 | 0.001536729 |
| BC080695 | 0.019547127 | 0.11561123 |
| BC147527 | 0.008057542 | 0.05154259 |
| Bcat1 | -0.02580247 | 0.009828978 |
| Bdnf | 0.017248236 | 0.022126278 |
| Bend7 | 0.012524447 | 0.000365108 |
| Bex1 | 0.052851204 | -0.011966423 |
| Bhlhe40 | 0.05307626 | -0.009928111 |

| | | |
|---|---|---|
| Bmp4 | -0.01896895 | -0.000544794 |
| Bnip3 | -0.014154435 | 0.009338872 |
| Bst2 | 0.03898057 | -0.027949888 |
| Btg1 | 0.021517655 | 0.029884223 |
| Btg2 | 0.03011527 | 0.036223702 |
| C130073F10Rik | -0.000534728 | 0.021176014 |
| C1d | 0.016326826 | 0.036491755 |
| C2cd4b | 0.004218822 | 0.027001975 |
| C730034F03Rik | 0.004175168 | 0.002167788 |
| C77370 | -0.010585178 | 3.70E-05 |
| Cacybp | -0.043136965 | -0.014030958 |
| Calb2 | -0.006771362 | 0.003902945 |
| Calcoco2 | -0.02539668 | 0.06307138 |
| Cald1 | 0.089361586 | 0.009736018 |
| Calm2 | 0.00668301 | -0.017183369 |
| Calml4 | 0.002918844 | -0.000479648 |
| Calr | 0.005403424 | 0.00282461 |
| Camk2d | 0.03708878 | -0.001303272 |
| Capg | 0.03781928 | -0.001552605 |
| Car2 | 0.08335073 | -0.056978278 |
| Car3 | 0.030876499 | -0.021049932 |
| Car4 | 0.06417472 | -0.012108661 |
| Carhsp1 | 0.023348901 | -0.009240522 |
| Catip | -0.024883859 | 0.001965289 |
| Cbr3 | 0.015500189 | -0.006089454 |
| Cbx3 | 0.000847163 | -0.026430955 |
| Ccnb1 | -0.003339117 | 0.008294192 |
| Ccnd1 | -0.013487862 | -0.02121775 |
| Ccnd2 | -0.003697974 | -0.018208168 |
| Ccnd3 | -0.03714356 | -0.00068462 |
| Ccne1 | -0.016595013 | -0.037401453 |
| Ccne2 | 0.014058882 | 0.002051401 |
| Ccng1 | 0.035949346 | 0.022919875 |
| Ccno | 0.010191255 | 0.04088658 |
| Ccser1 | 0.018254109 | 0.11549505 |
| Cd24a | 0.05137069 | 0.003102236 |
| Cd37 | -0.020764517 | 0.002678508 |
| Cd63 | 0.052259374 | 0.009962568 |

| | | |
|---|---|---|
| Cd74 | -0.010120583 | 0.012792378 |
| Cd81 | 0.035443403 | 0.016374405 |
| Cd9 | 0.002746664 | 0.01467389 |
| Cdc14a | 0.009268099 | 0.003439219 |
| Cdc20 | 0.003967631 | 0.000153341 |
| Cdc42ep3 | -0.000118792 | 0.038772427 |
| Cdc5l | -0.048449848 | 0.031973846 |
| Cdk6 | 0.032568462 | -0.023529418 |
| Cdkn1a | 0.0377119 | 0.013237253 |
| Cdkn1c | 0.021562133 | -0.015149301 |
| Cdkn2a | 0.028233293 | 0.004382915 |
| Cdv3 | 0.025066044 | -0.018136246 |
| Cdx2 | 0.034287825 | -0.006079288 |
| Cebpb | 0.037028495 | -0.002781794 |
| Cenpf | -0.00696432 | 0.008487805 |
| Cenpm | -0.04482287 | 0.028162606 |
| Cep57l1 | -0.005306352 | 0.04477432 |
| Chac1 | 0.007548773 | -0.014288743 |
| Chchd10 | -0.09330636 | 0.001045552 |
| Cidea | 0.002054567 | 0.04526647 |
| Cited1 | 0.04203834 | 0.010171789 |
| Cited2 | 0.055299006 | -0.003543699 |
| Cited4 | 0.017075216 | 0.007238478 |
| Ckb | -0.002592469 | 0.025500357 |
| Ckmt1 | 0.006780433 | 0.011775426 |
| Clcn5 | 0.03525746 | 0.020798948 |
| Cldn3 | 0.022947256 | -0.004431147 |
| Cldn4 | 0.040846772 | -0.003986615 |
| Cldn6 | 0.08158767 | -0.014180548 |
| Cldn7 | 0.073592074 | -0.01758181 |
| Clic1 | 0.06604909 | 0.003896484 |
| Clic4 | 0.013671224 | 0.003047593 |
| Clp1 | 0.003886337 | 0.050020095 |
| Cltb | -0.029335916 | -0.002472848 |
| Clu | 0.032374803 | 0.00506921 |
| Cmtm7 | 0.004018967 | -0.005194415 |
| Cnn1 | 0.061098456 | -0.01634129 |
| Cnn2 | 0.0597088 | -0.02715121 |

| | | |
|---|---|---|
| Cntfr | 0.027163005 | -0.02068379 |
| Coch | 0.011784329 | 0.027236966 |
| Col4a1 | 0.025383873 | 0.015150304 |
| Col4a2 | 0.029917192 | 0.008939553 |
| Commd3 | 0.026812669 | 0.021287989 |
| Comt | -0.016400581 | 0.006852488 |
| Cox6b2 | -0.013385477 | 0.039687626 |
| Cpsf4l | -0.014637545 | -0.005513027 |
| Crabp2 | 0.017374938 | -0.002946168 |
| Crct1 | 0.012285536 | 0.01212075 |
| Crim1 | 0.01852431 | 0.006244736 |
| Crip2 | 0.04015023 | -0.022849206 |
| Crxos | 0.001357752 | 0.04758899 |
| Cryab | 0.058373623 | 0.006871028 |
| Csrp1 | 0.08471262 | -0.0098939 |
| Cst3 | 0.035102163 | 0.010376593 |
| Cstb | 0.045886215 | 0.009141898 |
| Ctgf | 0.06567706 | -0.000594327 |
| Cth | -0.001842269 | 0.00329518 |
| Ctnnal1 | -0.016199414 | -0.007401111 |
| Ctr9 | 0.027460394 | 0.04338847 |
| Ctsa | 0.007530842 | 0.004168803 |
| Ctsb | 0.017083723 | 0.010107459 |
| Ctsd | 0.01733233 | 0.007856975 |
| Ctsl | 0.061471473 | 0.006524167 |
| Cul9 | -0.005204707 | 0.001919602 |
| Cwc22 | 0.008396202 | 0.056778036 |
| Cxcl16 | 0.060715612 | -0.003446068 |
| Cyba | 0.048274573 | -0.002684555 |
| Cyp26a1 | 0.013697446 | -0.002336123 |
| Cyr61 | 0.086159274 | -0.011615293 |
| D630023F18Rik | -0.0047449 | 0.0163669 |
| Dab2 | 0.031213526 | 0.0045787 |
| Dazl | 0.004006715 | 0.037785873 |
| Dcaf12l1 | 0.021828411 | 0.013338806 |
| Dcaf17 | 0.005013189 | 0.002799126 |
| Dcbld1 | -0.012647767 | 0.04350858 |
| Dcxr | 0.021073135 | 0.018912673 |

| | | |
|---|---|---|
| Ddah1 | 0.03765621 | -0.020219272 |
| Ddah2 | -0.003038738 | 0.015661106 |
| Ddit3 | 0.023030674 | 0.00745412 |
| Ddit4 | 0.038177192 | 0.03966237 |
| Ddit4l | 0.028028483 | 0.065412395 |
| Ddr2 | -0.000113222 | 0.03403952 |
| Derl3 | -0.003419998 | 0.010177033 |
| Dhx16 | -0.042943772 | 0.000432337 |
| Dkkl1 | 0.004654707 | 0.005099509 |
| Dlg1 | -0.002645243 | 0.007913143 |
| Dll1 | 0.003777089 | 0.007062126 |
| Dmkn | 0.01862967 | 0.006954388 |
| Dnajb14 | 0.007831642 | 0.03626714 |
| Dnajb9 | 0.009901268 | 0.046734184 |
| Dnajc21 | -0.044168252 | -0.009726992 |
| Dnmt3a | 0.000708433 | -0.01766263 |
| Dnmt3b | 0.034354787 | -0.047587495 |
| Dnmt3l | -0.020602236 | -0.020090772 |
| Dok2 | 0.039054066 | -0.017521095 |
| Dpp7 | 0.002323445 | 0.007300701 |
| Dppa2 | -0.023127988 | 0.05652287 |
| Dppa3 | 0.007235021 | 0.01017705 |
| Dppa5a | -0.08626957 | 0.041418824 |
| Dpys | -0.021740822 | -0.003951479 |
| Dqx1 | -0.016957076 | 0.022540657 |
| Dsp | 0.043254197 | 0.000568998 |
| Dst | 0.024845371 | 0.048654616 |
| Dstn | 0.039652593 | -0.03151753 |
| Dusp1 | 0.049871363 | 0.008013587 |
| Dusp14 | 0.0334446 | -0.002270326 |
| Dusp4 | 0.054571792 | -0.010000737 |
| Dusp5 | 0.043761633 | 0.00626048 |
| Dusp6 | 0.051718485 | -0.023772363 |
| Dusp9 | 0.01181286 | 0.018920012 |
| Duxf3 | 0.011258294 | 0.07037724 |
| Dynll1 | -0.00620102 | 0.011316283 |
| E130012A19Rik | -0.076676786 | 0.000726886 |
| E2f5 | 0.00865022 | -0.004996888 |

| | | |
|---|---|---|
| Ebf1 | 0.010503617 | -0.008187851 |
| Echdc2 | -0.000268897 | 0.025201052 |
| Echdc3 | 0.012568678 | 0.05705047 |
| Edn1 | 0.011678997 | 0.004059735 |
| Edn2 | -0.003892543 | -0.003289369 |
| Eef1a1 | 0.021844989 | 0.05794018 |
| Eef2 | 0.029955793 | 0.017947892 |
| Efhc2 | 0.00206725 | 0.002763808 |
| Efna3 | 0.010761168 | 0.007722148 |
| Efnb1 | 0.030413523 | -0.001442606 |
| Egr1 | -0.000749228 | -0.007151757 |
| Egr4 | -0.000124314 | -0.00164326 |
| Eif2s2 | -0.062068503 | 0.0475361 |
| Eif4a1 | 0.000289934 | -0.004212175 |
| Elf3 | 0.015611121 | 0.011153074 |
| Emb | 0.019994577 | -0.030172413 |
| Emilin1 | -0.016405562 | 0.008160953 |
| Emp2 | 0.023016015 | 0.002261004 |
| Emp3 | 0.029896982 | 0.003426032 |
| En1 | -0.000505638 | -0.0062662 |
| Enah | -0.05129853 | -0.010647598 |
| Endov | 0.000849611 | -0.003806706 |
| Eno1 | 0.023922602 | -0.03532054 |
| Enpp2 | 0.018391622 | -0.013291356 |
| Epcam | 0.050913747 | -0.02809695 |
| Epha2 | 0.027004072 | -0.021093363 |
| Eps8l2 | 0.028736366 | 0.008347368 |
| Ercc4 | 0.012984146 | 0.10032243 |
| Erdr1 | 0.030274814 | -0.027269442 |
| Esco2 | -0.037606016 | 0.023749426 |
| Esx1 | 0.018991467 | 0.007406469 |
| Eva1a | 0.003622026 | 0.009416269 |
| F2r | 0.039685924 | -0.014354883 |
| F2rl1 | 0.018891457 | -0.022135587 |
| F3 | 0.034614768 | -0.004506289 |
| Fabp3 | 0.023035884 | 0.014178606 |
| Fabp5 | 0.05758098 | -0.021719424 |
| Fam124a | 0.012689043 | 0.05592422 |

| | | |
|---|---|---|
| Fam134b | 0.016535224 | 0.004645391 |
| Fam159b | 0.006285528 | 0.000941234 |
| Fam162a | -0.009869767 | 0.003922158 |
| Fam25c | -0.04846094 | 0.021403132 |
| Fam60a | 0.005196818 | -0.056118388 |
| Fam83a | -0.000325609 | 0.001906392 |
| Fblim1 | -0.04170016 | -0.022100853 |
| Fbp2 | 0.005118796 | 0.02218248 |
| Fbxo15 | -0.010482747 | 0.07644365 |
| Fgf1 | 0.007225305 | 0.05742421 |
| Fgf15 | 9.95E-05 | -0.005512751 |
| Fgf17 | -0.012837992 | -0.014580964 |
| Fgf4 | -0.036527693 | 0.007016205 |
| Fgf5 | 0.023556119 | -0.006828413 |
| Fgf8 | 0.007930755 | -0.003702276 |
| Fgfbp1 | 0.07556955 | -0.018307226 |
| Fhl1 | 0.026304467 | -0.006946371 |
| Fhl2 | 0.048406772 | -0.003450101 |
| Fkbp6 | -0.011712599 | 0.018409675 |
| Flrt3 | 0.011473482 | 0.002489002 |
| Fmr1nb | 0.002148234 | 0.034822974 |
| Folr1 | 0.021634636 | -0.005113195 |
| Fos | 0.01829026 | 0.006016003 |
| Foxo3 | 0.000551056 | 0.017125327 |
| Fst | 0.03634429 | -0.008764674 |
| Fth1 | -0.000284873 | -0.054010097 |
| Fthl17a | 0.012782832 | 0.008670297 |
| Fthl17c | 0.002489224 | 0.037902582 |
| Fuca2 | 0.016489467 | -0.00165122 |
| Fundc1 | 0.018053822 | 0.051176306 |
| Fxyd6 | -0.002643741 | -0.008005482 |
| G0s2 | -0.003433945 | 0.000814898 |
| Gabarapl2 | -0.07046195 | 0.07336173 |
| Gadd45a | 0.013230865 | -0.020593867 |
| Gadd45b | 0.034734026 | 0.011001428 |
| Gadd45g | 0.016840952 | -0.005709146 |
| Gap43 | 0.028587352 | -0.018545061 |
| Gata2 | 0.021185473 | 0.01101415 |

| | | |
|---|---|---|
| Gata3 | 0.020810654 | 0.012187039 |
| Gbp2 | 0.047714606 | -0.010505687 |
| Gbp2b | 0.0413234 | -0.008629489 |
| Gbx2 | -0.002578741 | -0.030037615 |
| Gchfr | 0.023111401 | -0.00524553 |
| Gfod2 | -0.002669147 | 0.000774122 |
| Gja1 | 0.03729728 | -0.022795988 |
| Gjb3 | 0.000899088 | 0.001832224 |
| Gjb5 | -0.016587796 | 0.004303272 |
| Glipr2 | 0.03878476 | -0.010613338 |
| Glod5 | -0.01198676 | -0.008469946 |
| Glrx | 0.018930193 | 0.056158848 |
| Glrx2 | 0.022932766 | 0.07689187 |
| Glul | -0.007651395 | 0.002570058 |
| Gm10116 | 0.009434057 | -0.02134013 |
| Gm11238 | 0.00871849 | 0.05774779 |
| Gm12794 | 0.011987051 | 0.06922997 |
| Gm13075 | -0.026613394 | 0.001503398 |
| Gm13119 | 0.01652407 | 0.10431137 |
| Gm13145 | -0.017275494 | 0.023041628 |
| Gm13225 | -0.035756707 | 0.015297518 |
| Gm15879 | -0.000230966 | 0.003720969 |
| Gm15915 | -0.008747464 | 0.003994174 |
| Gm16233 | 0.056657385 | -0.023414169 |
| Gm2016 | 0.019900931 | 0.13346794 |
| Gm2035 | 0.019177075 | 0.12121134 |
| Gm2046 | 0.017295599 | 0.10846861 |
| Gm2056 | 0.019292343 | 0.12318259 |
| Gm20767 | 0.015838897 | 0.09309257 |
| Gm21761 | 0.028489914 | 0.16648936 |
| Gm21818 | 0.020783337 | 0.12639622 |
| Gm26710 | -0.001675538 | -0.002017949 |
| Gm26737 | -0.002618892 | 0.010177097 |
| Gm26782 | -0.002579627 | 0.000771172 |
| Gm26870 | -0.006236042 | 0.09962724 |
| Gm26917 | -0.024129013 | 0.023420332 |
| Gm27167 | 0.010649298 | 0.07257229 |
| Gm29666 | -0.009751109 | 0.005910335 |

| | | |
|---|---|---|
| Gm33466 | 0.010937145 | 0.066742495 |
| Gm36266 | -0.002551836 | -0.000720445 |
| Gm37305 | -0.005015675 | 0.001437683 |
| Gm4027 | 0.019480506 | 0.120306306 |
| Gm42418 | -0.008831644 | 0.015085356 |
| Gm42637 | 0.028787496 | 0.003784247 |
| Gm428 | 0.008940176 | 0.059008334 |
| Gm4340 | 0.015008505 | 0.09167015 |
| Gm43409 | -0.006639358 | 0.00504022 |
| Gm4349 | 0.000587295 | 0.032809667 |
| Gm5039 | 0.018662311 | 0.11976821 |
| Gm5662 | 0.026117979 | 0.17187935 |
| Gm7120 | -0.001746533 | 0.017229041 |
| Gm7325 | -0.056749877 | -0.002618413 |
| Gm8300 | 0.02716212 | 0.17303982 |
| Gm8994 | 0.02206793 | 0.13588205 |
| Gm9112 | -0.000240859 | 0.005932308 |
| Gmcl1 | -0.004257055 | 0.002705255 |
| Gnas | 0.049456082 | -0.00803878 |
| Golga3 | -0.013204619 | 0.021294672 |
| Gpd1l | -0.004854105 | -0.001267369 |
| Gprc5a | 0.030196998 | 0.001127704 |
| Gprc5c | 0.022903036 | 0.010446035 |
| Gpx3 | 0.025270222 | 0.006060973 |
| Gpx4 | -0.044192065 | 0.04487611 |
| Grb10 | 0.029814072 | -0.023128377 |
| Grn | 0.007838422 | 0.02726066 |
| Gsn | 0.051334053 | 0.001813137 |
| Gsta4 | -0.034142185 | -0.009995136 |
| Gstm1 | 0.008087546 | -0.000892051 |
| Gsto1 | 0.034149937 | -0.0021975 |
| Gtsf1l | -0.055739395 | -0.002540977 |
| Guca1a | 0.004055876 | 0.055705376 |
| H19 | 0.026705291 | 0.005265772 |
| H1f0 | 0.01245636 | 0.023837402 |
| H2-D1 | 0.021119803 | 0.00669263 |
| H2-T23 | -0.00580988 | 0.005367006 |
| H2afx | -0.0426532 | 0.023524415 |

| | | |
|---|---|---|
| H2afy | 0.006108494 | 0.002557848 |
| H3f3b | 0.03499717 | 0.007577031 |
| Hand1 | 0.02656063 | 0.004488193 |
| Hat1 | -0.012885354 | -0.000537764 |
| Hbegf | 0.025257757 | -0.002340328 |
| Hck | -0.021693593 | -0.001472522 |
| Hdc | 0.006842143 | 0.03282815 |
| Herpud1 | 0.008867953 | -0.000240387 |
| Hes1 | 0.0073727 | 0.002403795 |
| Hes6 | 0.030074546 | -0.016650561 |
| Hexa | 0.01706672 | 0.009382461 |
| Hexb | -0.00768192 | 0.019732006 |
| Hist1h1a | -0.006710325 | 0.002509512 |
| Hist1h1b | -0.010880776 | -0.018049197 |
| Hist1h1e | 0.014197714 | -0.014656152 |
| Hist1h2aa | 0.001198959 | -0.001707079 |
| Hist1h2ae | -0.004707729 | -0.004079382 |
| Hist1h2ap | -0.004071913 | -0.00509812 |
| Hist1h2bc | -0.003143966 | 0.035367057 |
| Hist1h3c | 0.005970402 | 0.03331994 |
| Hist1h3d | 0.001558883 | 0.023724755 |
| Hist1h4i | 0.008683445 | 0.06782613 |
| Hist1h4j | 0.003598591 | 0.023746489 |
| Hist2h3c1 | 0.00390085 | 0.029336303 |
| Hist3h2a | -0.000197058 | 0.005012439 |
| Hist3h2ba | 0.013614152 | 0.027875219 |
| Hmces | -0.06936463 | -0.020522393 |
| Hmga2 | 0.055883076 | -0.00339953 |
| Hmgb3 | 0.026266526 | -0.008529956 |
| Hmgcr | 0.016960127 | -0.007389274 |
| Hmgn3 | 0.047155555 | 0.052114233 |
| Hmgn5 | 0.03150653 | 0.015951917 |
| Hmmr | -0.009572479 | 0.032740057 |
| Hmox1 | 0.024818212 | -0.022696791 |
| Hnrnpa1 | 0.03001399 | -0.02950481 |
| Hormad1 | -0.009406035 | 0.011291593 |
| Hoxa7 | 0.008204411 | 0.000837694 |
| Hpcal1 | 0.02258307 | 0.000602338 |

| | | |
|---|---|---|
| Hsd17b14 | -0.10991553 | -0.026478708 |
| Hsf2 | 0.005301386 | -0.003861517 |
| Hsp90aa1 | 0.007263745 | 0.01477534 |
| Hsp90ab1 | -0.010708533 | 0.04429141 |
| Hsp90b1 | 0.014367413 | -0.005401909 |
| Hspa1a | 0.011883227 | 0.013529325 |
| Hspa5 | 0.019750461 | -0.004727841 |
| Hspa8 | 0.008355865 | 0.030436693 |
| Hspb1 | -0.049969357 | 0.04167702 |
| Hspb2 | 0.028570263 | 0.004832401 |
| Htra1 | 0.012080606 | 0.004158956 |
| Icam1 | -0.010909056 | -0.004315316 |
| Id1 | 0.02968073 | -0.001653581 |
| Id2 | 0.057427086 | 0.000848706 |
| Id3 | 0.01877864 | 0.025810366 |
| Ier3 | 0.020002209 | -0.003742611 |
| Ier5 | 0.033289887 | -0.001175506 |
| Ifi30 | 0.014150376 | -0.000506245 |
| Ifitm1 | -0.022298144 | -0.015004985 |
| Ifitm2 | -0.083204456 | -0.007295612 |
| Ifitm3 | 0.023948621 | -0.02431863 |
| Ifrd1 | 0.04163054 | -0.01607971 |
| Igf2 | 0.05488462 | 0.002633388 |
| Igfbp2 | 0.00736198 | 0.034500096 |
| Igfbp3 | 0.05463665 | 0.034464307 |
| Igfbp4 | 0.025155067 | -0.009441111 |
| Igsf23 | -0.019173983 | 0.005373285 |
| Il23a | 0.001476667 | -0.001180959 |
| Impact | 0.03791328 | -0.021357508 |
| Ina | -0.007049399 | -0.037153456 |
| Insig1 | 0.007154108 | -0.001100667 |
| Irf1 | 0.02699284 | -0.011912086 |
| Irgm1 | 0.013979862 | -0.003943445 |
| Itga3 | 0.056282826 | -0.016729996 |
| Itgb1 | 0.049246784 | -0.02181563 |
| Itm2a | 0.01544533 | 0.012279341 |
| Jam2 | -0.06360303 | 0.050534934 |
| Jarid2 | -0.037127305 | -0.013621271 |

| | | |
|---|---|---|
| Jun | 0.059227612 | 0.021227775 |
| Junb | 0.028771855 | -0.022570617 |
| Jund | 0.04185853 | -0.013093876 |
| Kcnk1 | 0.052004386 | -0.020354835 |
| Kctd12 | 0.010844702 | 0.016986847 |
| Kdelr3 | -0.012378245 | 0.003609421 |
| Kdm5b | -0.047196697 | 0.03207779 |
| Kdm6b | 0.000870149 | -0.004686068 |
| Khdc3 | -0.039540254 | 0.016528096 |
| Kif1a | 0.043843795 | -0.010401619 |
| Kif21a | 0.020778624 | -0.010200771 |
| Klf2 | -0.08013138 | 0.016469633 |
| Klf4 | -0.059980217 | 0.018691417 |
| Klf5 | -0.009370692 | -0.00288512 |
| Klf6 | 0.06657653 | -0.007460462 |
| Klf9 | -0.01590595 | 0.021249514 |
| Klhl13 | 0.029314125 | 0.020981133 |
| Kpna7 | -0.007447249 | 0.004292301 |
| Kras | 0.049295247 | -0.0309612 |
| Krt17 | -0.000173826 | -0.005908305 |
| Krt18 | 0.10321411 | 0.000387038 |
| Krt19 | 0.07333735 | -0.004182681 |
| Krt42 | -0.029897258 | 0.000875479 |
| Krt7 | 0.04993038 | 0.002116194 |
| Krt8 | 0.10330999 | 0.003748034 |
| L1td1 | -0.026675548 | -0.04371335 |
| Lama1 | -0.008177145 | 0.001823619 |
| Lamb1 | 0.026096877 | -0.003363888 |
| Ldha | -0.000971542 | -0.047040686 |
| Ldhb | -0.01845672 | 0.021210525 |
| Ldoc1 | 0.014645477 | 0.00014748 |
| Lef1 | 0.004280104 | -0.02849528 |
| Lefty1 | 0.05940725 | -0.025700396 |
| Lefty2 | 0.04773168 | -0.019853717 |
| Lepr | -0.011732918 | 0.017883783 |
| Lgals1 | 0.0744673 | -0.016787278 |
| Lgals3 | -0.030066986 | 0.021918105 |
| Lgals4 | 0.020937644 | 0.104770355 |

| | | |
|---|---|---|
| Lgmn | 0.011774233 | 0.00695827 |
| Lgr4 | 0.005211347 | -0.011665749 |
| Limch1 | 0.017973818 | 0.08596835 |
| Limk2 | -0.003005705 | 0.006138702 |
| Lmo4 | 0.03952487 | -0.000920154 |
| Lmo7 | -0.000393714 | 0.010008211 |
| Lmx1a | 0.010885065 | 0.058181003 |
| Lncenc1 | -0.043541767 | 0.004422141 |
| Lonp2 | 0.020628309 | 0.08903978 |
| Lor | 0.029897433 | 0.001895448 |
| Loxl2 | -0.009430684 | 0.0063161 |
| Lpar1 | -0.011263983 | 0.022718772 |
| Lpar4 | 0.028974632 | -0.011288257 |
| Lpar6 | 0.042156607 | -0.001224014 |
| Lpl | 0.01072217 | 0.012571201 |
| Lpp | 0.048111763 | -0.009801238 |
| Lrp2 | 0.026291138 | -0.006818071 |
| Lrpap1 | 0.01796728 | 0.025242886 |
| Lrrc8a | 0.025152612 | -0.006021878 |
| Lrrn4 | 0.030134331 | 0.021347338 |
| Ltb | 0.044556923 | -0.009757914 |
| Ly6a | 0.000215918 | 0.019536823 |
| Ly6g6e | -0.014162073 | 0.014379009 |
| Ly6k | -0.001906297 | 0.02489482 |
| Malat1 | 0.057314955 | -0.03672508 |
| Manba | -0.051747307 | 0.008637489 |
| Manf | 0.010473885 | -0.000702933 |
| Map1b | 0.014175134 | -0.011072679 |
| Mapt | -0.031076511 | 0.007832727 |
| Marcks | 0.036807302 | -0.021691866 |
| Marcksl1 | 0.022543961 | -0.031025883 |
| Marf1 | 0.000749115 | -0.002467677 |
| Marveld1 | 0.01210741 | 0.047261484 |
| Mbd5 | 0.006405154 | 0.030411066 |
| Mbnl3 | 0.024516828 | 0.000193047 |
| Mdk | 0.031358805 | -0.016537867 |
| Mdm2 | 0.030469349 | 0.018095024 |
| Mecom | 0.015892126 | 0.07152331 |

| | | |
|---|---|---|
| Meg3 | 0.019786512 | -0.057354063 |
| Mest | 0.045289718 | -0.026361784 |
| Mfge8 | 0.005932598 | 0.022693532 |
| Mfsd1 | 0.016010303 | 0.01473751 |
| Mgll | 0.01750019 | 0.08567064 |
| Micu1 | 0.009235302 | 0.034322266 |
| Mif | 0.007847809 | -0.07539473 |
| Mis18bp1 | -0.012069989 | 0.046463285 |
| Mkrn1 | -0.073402606 | 0.011514884 |
| Mme | 0.012132841 | 0.013886697 |
| Mobp | -0.010899177 | 0.011042444 |
| Morc1 | -0.061956212 | 0.008285911 |
| Morc4 | 0.027526155 | 0.0080546 |
| Mpzl1 | 0.026916765 | 0.038218185 |
| Mras | -0.05014885 | -0.00091585 |
| Mreg | -0.015978057 | 0.06502478 |
| Mrpl9 | 0.003990512 | -0.002522284 |
| Ms4a10 | 0.001597203 | 0.012653625 |
| Msc | -0.035349235 | 0.000809904 |
| Msmo1 | 0.008451441 | -0.007711607 |
| Msx1 | 0.02230118 | 0.004735145 |
| Msx2 | 0.01838304 | 0.028504135 |
| mt-Atp6 | -0.060005236 | -0.00461201 |
| mt-Co1 | -0.04622106 | 0.017192287 |
| mt-Co2 | -0.03966278 | 0.007804764 |
| mt-Co3 | -0.06383781 | 0.007576443 |
| mt-Cytb | -0.05752422 | -0.00251884 |
| mt-Nd1 | -0.060112312 | -0.003382622 |
| mt-Nd2 | -0.035542604 | 0.004146815 |
| mt-Nd4 | -0.04583455 | 0.013814743 |
| Mt1 | 0.005800775 | -0.019519608 |
| Mt2 | -0.003861539 | -0.009523704 |
| Mtf2 | -0.052479256 | 0.045621753 |
| Mthfd2 | 0.003420014 | -0.020894522 |
| Muc3 | -0.005166563 | 0.001622851 |
| Myc | 0.007466979 | -0.019138174 |
| Mycn | 0.014987506 | -0.04518277 |
| Myef2 | 0.010676716 | -0.018849714 |

| | | |
|---|---|---|
| Myh13 | -0.006115272 | 0.023224454 |
| Myl4 | -0.005642884 | 0.011814928 |
| Myl6 | 0.03668839 | 0.001788623 |
| Myl9 | 0.045190603 | -0.019326176 |
| Mylpf | -0.08297398 | 0.04381212 |
| Myo1f | -0.03700731 | 0.003271781 |
| Nab2 | 0.002168507 | -0.004773438 |
| Nanog | -0.040753033 | 0.032297507 |
| Napsa | -0.014370701 | -0.006378022 |
| Ncoa3 | -0.02217006 | -0.00133258 |
| Ndrg1 | 0.009102671 | 0.005070039 |
| Ndufa4l2 | -0.005779321 | 0.009633251 |
| Neat1 | 0.01802083 | -0.006658479 |
| Nefh | -0.014870527 | 0.003422987 |
| Nefl | 0.015276651 | -0.009227906 |
| Nefm | 0.020993007 | -0.002136491 |
| Nelfa | 0.010052733 | 0.07832915 |
| Nes | 0.064570814 | -0.016530758 |
| Neurod1 | 0.006598843 | 0.00090453 |
| Neurog3 | -0.005914768 | -0.000706828 |
| Nexn | 0.010470851 | -0.008434624 |
| Nfatc2ip | -0.052524477 | 0.026617693 |
| Nfkbia | 0.012581361 | 0.032037217 |
| Ngb | 2.21E-05 | -0.00129646 |
| Ngfrap1 | 0.024133451 | 0.007734808 |
| Nip7 | -0.047540464 | -0.013668732 |
| Nkx2-9 | 0.001946361 | 0.007675432 |
| Nkx6-2 | -0.004339927 | -0.001939896 |
| Nmrk1 | 7.99E-05 | 0.006283409 |
| Nnat | 0.049553543 | -0.01899711 |
| Noct | 0.000807107 | -0.003735249 |
| Nodal | -0.002352786 | 0.011091408 |
| Nop56 | 0.001066659 | 0.00089646 |
| Notch2 | 0.021051958 | -0.007071473 |
| Nptxr | 0.007347569 | -0.001191676 |
| Npy | 0.001377303 | 0.020490708 |
| Nr0b1 | -0.06055917 | -0.002848119 |
| Nrk | 0.023927797 | 0.004806237 |

| | | |
|---|---|---|
| Nrp2 | 0.015762761 | -0.007330699 |
| Nudt4 | 0.00515578 | 0.02170633 |
| Nup62cl | 0.018903123 | 0.013518163 |
| Nupr1 | -0.004878024 | -0.004618408 |
| Ooep | -0.044575837 | 0.017879413 |
| Os9 | 0.011851691 | 0.05213031 |
| Otx2 | -0.003979108 | -0.009220701 |
| P2rx4 | 0.007916019 | 0.002534957 |
| P2rx7 | 0.008146117 | 0.005876616 |
| P3h3 | -0.006478303 | 0.007346639 |
| P3h4 | -0.00092829 | -0.024164522 |
| P4ha2 | 0.026330207 | 0.050529126 |
| Parm1 | 0.00611063 | 0.010557724 |
| Parp1 | -0.051629543 | 0.009794451 |
| Pawr | 0.036791798 | -0.008251214 |
| Pax6 | -0.013641947 | -0.013556893 |
| Pcgf5 | 0.039054286 | 0.013535514 |
| Pdcd4 | 0.000435608 | 0.013173285 |
| Pdgfa | -0.018076453 | 0.011351427 |
| Pdgfrl | 0.020569652 | 0.08402794 |
| Pdia3 | 0.022405997 | 0.007023453 |
| Pdia6 | 0.015307296 | -0.004670825 |
| Pdlim3 | 0.01799686 | 0.035846457 |
| Pdlim4 | 0.008709206 | 0.006746862 |
| Pdlim7 | 0.07853086 | -0.014697516 |
| Pdpn | 0.021638455 | -0.000112674 |
| Peg10 | 0.05086583 | -0.015998336 |
| Peg3 | 0.024057368 | -0.00206 |
| Perp | 0.08387126 | -0.00742062 |
| Pfas | -0.004540946 | -0.016702062 |
| Pfkp | -0.033138867 | 0.01696471 |
| Pfn1 | 0.03361296 | -0.07398718 |
| Pga5 | -0.010424935 | 0.008379809 |
| Phgdh | 0.03325548 | -0.016655719 |
| Phkg1 | -0.008103333 | 0.004237564 |
| Phlda1 | 0.055025104 | 0.003274372 |
| Phlda2 | 0.026554054 | -0.026392858 |
| Phlda3 | 0.04163544 | 0.007687937 |

| | | |
|---|---|---|
| Phldb2 | 0.04849776 | 0.019127484 |
| Pim1 | 0.04880737 | -0.003834302 |
| Pim2 | 0.05631185 | -0.040253673 |
| Pitx2 | 0.051216763 | -0.013198326 |
| Pkp2 | 0.003561762 | -0.003274379 |
| Plac8 | 0.01615424 | -0.006697566 |
| Platr3 | -0.041993182 | 0.019080129 |
| Platr31 | 0.020313788 | 0.12293335 |
| Plaur | 0.062150825 | -0.016355803 |
| Plet1 | 0.021496253 | 0.00272047 |
| Plin2 | 0.067612395 | 0.017473547 |
| Plk2 | 0.039671298 | 0.023353897 |
| Plk3 | 0.013478932 | 0.005904021 |
| Plk4 | -0.03257671 | 0.007059331 |
| Plod2 | 0.022805115 | 0.008123179 |
| Pmaip1 | 0.03217391 | 0.03838537 |
| Pmm1 | -0.00285713 | -0.012517266 |
| Pmp22 | 0.04252212 | -0.002578012 |
| Pou3f1 | 0.0886468 | -0.03891264 |
| Ppdpf | -0.041057065 | 0.001050807 |
| Ppig | 0.004191735 | 0.03534577 |
| Pramel7 | 0.003946993 | 0.030915057 |
| Prdx1 | -0.0015816 | -0.02855898 |
| Prdx6 | 0.040108826 | -0.005933631 |
| Prex2 | 0.018188821 | 0.07823154 |
| Prph | 0.024338495 | -0.003125917 |
| Prps1 | 0.002218143 | 0.035233956 |
| Prr13 | -0.005223257 | -0.002529654 |
| Prrc1 | -0.024680434 | 0.028655814 |
| Prss8 | 0.039840695 | -0.007705758 |
| Prtg | 0.06548481 | -0.02052452 |
| Psap | 0.017358545 | 0.025149843 |
| Psmb8 | 0.03527508 | -0.010623768 |
| Psors1c2 | 0.016894786 | -0.007919516 |
| Psrc1 | 0.002479071 | -0.000257843 |
| Ptch1 | -0.001731897 | -0.012943762 |
| Ptcra | -0.013300842 | 0.000844152 |
| Ptges | 0.009105714 | 0.003957763 |

| | | |
|---|---|---|
| Ptgr1 | 0.017969768 | 0.018110372 |
| Pth1r | 0.01548752 | 0.000931845 |
| Ptp4a2 | 0.02229632 | -0.01127208 |
| Ptp4a3 | 0.004481949 | 0.008163252 |
| Pttg1 | 0.031276498 | 0.021524634 |
| Pxdc1 | 0.046567295 | -0.002733822 |
| Pycr2 | 0.04551226 | -0.0328982 |
| Rab3il1 | 0.009223086 | 0.003722931 |
| Rap2b | 0.036107548 | -0.001991218 |
| Rasgrp2 | -0.03861917 | -0.026290286 |
| Rasl11a | 0.005770573 | 0.00970229 |
| Rassf1 | 0.029769907 | -0.009443105 |
| Rbm15 | 0.003283433 | -0.002860165 |
| Rbm47 | -0.000416183 | 0.004094035 |
| Rbms1 | 0.06797447 | -0.016751237 |
| Rbmx | 0.02304335 | 0.031674378 |
| Rbp1 | 0.016458854 | 0.007730162 |
| Rbpj | -0.03953207 | 0.003767239 |
| Rcsd1 | 0.00263533 | 0.0234671 |
| Rec114 | -0.008597365 | 0.019676246 |
| Reep5 | 0.06302246 | -0.001369909 |
| Renbp | 0.007957466 | 0.022281904 |
| Rest | -0.04838758 | -0.038601115 |
| Retn | -0.006249201 | 0.002707783 |
| Rgcc | 0.011056006 | 0.012279421 |
| Rhob | 0.031095633 | -0.002731184 |
| Rhox13 | -0.007728287 | 0.020375602 |
| Rhox5 | 0.004451726 | 0.013406365 |
| Rhox6 | 0.024675332 | 0.007412664 |
| Rhox9 | 0.02135351 | 0.008256732 |
| Rif1 | -0.027877264 | 0.011302024 |
| Rimklb | 0.006817012 | 0.026108313 |
| Rnf128 | 0.059317537 | -0.004290893 |
| Rnf168 | -0.004360375 | 0.001609128 |
| Rnf7 | -0.028881129 | 0.002017015 |
| RP23-4H17.3 | -0.036238898 | 0.01924155 |
| Rpl10l | -0.01659579 | 0.022288768 |
| Rpl22l1 | -0.017381923 | 0.013222387 |

| | | |
|---|---|---|
| Rpl39l | -0.056015678 | 0.03315723 |
| Rpl41 | -0.015711447 | -0.019872895 |
| Rpp25 | -0.050927423 | 0.02583105 |
| Rps18 | -0.027430108 | 0.037106704 |
| Rps19 | 0.012676132 | 0.026450522 |
| Rps23 | -0.019110046 | 0.04952674 |
| Rps27l | 0.004749517 | -0.016248133 |
| Rps4x | -0.012563138 | 0.045727815 |
| Rrad | 0.024700148 | -0.005643126 |
| Rrm2 | 0.004064 | -0.004658698 |
| Rrn3 | -0.017104777 | -0.007383612 |
| Rsph6a | 0.002253769 | 0.0047081 |
| Rsrp1 | 0.005601133 | 0.012154623 |
| Rtkn | -0.000861793 | 0.0146809 |
| S100a1 | 0.016439274 | 0.007270871 |
| S100a10 | 0.051903922 | 0.010155744 |
| S100a11 | 0.057441738 | 0.03803487 |
| S100a13 | 0.0205571 | 0.018000882 |
| S100a3 | 0.012969553 | -0.00112757 |
| S100a6 | 0.068982095 | -0.001198798 |
| Sall3 | -0.001860465 | -0.000654704 |
| Sall4 | -0.012040203 | 0.009051272 |
| Sat1 | 0.014477412 | 0.020402886 |
| Sbsn | 0.03346857 | 0.003151454 |
| Scamp1 | 0.02961219 | 0.09533558 |
| Scd2 | 0.007169057 | -0.001180985 |
| Scn1b | 0.02148244 | 0.005727404 |
| Sct | 0.031138703 | 0.019131266 |
| Sdc4 | 0.034923624 | 0.0194648 |
| Sdhaf3 | 0.003659992 | 0.0420364 |
| Sec24a | -0.003607738 | 0.007881408 |
| Sema6a | 0.035791382 | -0.011948902 |
| Sepp1 | 0.019565145 | -0.002290045 |
| Sep-01 | -0.076747864 | 0.004510315 |
| Sepw1 | 0.027275132 | -0.029971553 |
| Serpinb9b | 0.023421612 | -0.000467022 |
| Serpine2 | 0.028537424 | 0.018052759 |
| Serpinh1 | 0.08123845 | -0.008475008 |

| | | |
|---|---|---|
| Sfmbt2 | 0.037797254 | -0.006549982 |
| Sfn | 0.08270626 | -0.005199599 |
| Sfrp1 | -0.009837274 | -0.0031103 |
| Sgk1 | -0.021003142 | -0.010393793 |
| Shkbp1 | 0.008606517 | -0.0016724 |
| Sin3b | 0.02685117 | 0.014627228 |
| Six1 | 0.007049811 | -0.003473594 |
| Skil | 0.0375166 | -0.029608008 |
| Slamf1 | -0.01944768 | 0.013315115 |
| Slbp | 0.001693504 | -0.015306074 |
| Slc11a1 | -0.025275243 | 0.004640296 |
| Slc16a3 | 0.037926756 | -0.027918458 |
| Slc24a5 | 0.009730256 | -0.018596807 |
| Slc25a20 | -0.010425132 | 0.01832664 |
| Slc25a31 | 0.007528547 | 0.022165133 |
| Slc25a4 | 0.06122272 | 0.00475799 |
| Slc25a43 | 0.003131504 | 0.000501298 |
| Slc25a5 | -0.036147635 | 0.009848697 |
| Slc26a2 | 0.011719673 | 0.004974777 |
| Slc27a2 | 0.017212173 | 0.032033417 |
| Slc29a1 | -0.03954156 | -0.009067782 |
| Slc2a1 | 0.03597169 | -0.014579083 |
| Slc2a3 | 0.027532816 | 0.013592348 |
| Slc30a2 | 0.001976221 | 0.000553248 |
| Slc39a1 | 0.007990097 | -0.006193065 |
| Slc39a8 | 0.03822727 | -0.015664298 |
| Slc40a1 | 0.03052282 | -0.000360589 |
| Slc7a3 | -0.023246983 | -0.013905251 |
| Smad7 | 0.009136478 | -0.004067719 |
| Smagp | -0.00623281 | -0.03939039 |
| Smarca5 | -0.008255747 | 0.017914219 |
| Sms | -0.019401489 | -0.01808401 |
| Smtnl1 | -0.01521073 | 0.010908905 |
| Snai1 | 0.015372855 | 0.033827316 |
| Snhg12 | 0.001842588 | 0.001062996 |
| Snrpn | 0.018666046 | -0.028992375 |
| Snx20 | 0.002996625 | 0.005710765 |
| Socs2 | 0.006881579 | -0.03149664 |

| | | |
|---|---|---|
| Socs3 | -0.012961913 | 0.011435423 |
| Sod2 | -0.066615276 | 0.0067039 |
| Soga3 | -0.018426048 | -0.008850227 |
| Sord | 0.008766575 | 0.04990718 |
| Sox11 | 0.037886783 | -0.022879835 |
| Sox15 | -0.016951924 | 0.010293589 |
| Sox2 | -0.05155959 | -0.02045335 |
| Sox21 | -0.009492586 | -0.000448331 |
| Sox4 | 0.03250007 | -0.028612142 |
| Sp110 | 0.013626587 | 0.08309444 |
| Sp140 | 0.010037819 | 0.04448239 |
| Sp5 | 0.025646172 | -0.014795496 |
| Sparc | 0.034611516 | 0.01938762 |
| Spesp1 | 0.011669086 | 0.06768652 |
| Spic | -0.004022508 | 0.018942254 |
| Spink1 | -0.05046564 | 0.0025592 |
| Spp1 | -0.05136179 | 0.005936741 |
| Spry2 | -0.004363427 | -0.011807059 |
| Spry4 | -0.014325032 | -0.011115747 |
| Spty2d1 | -0.018075317 | 0.030452097 |
| Sqstm1 | 0.029067406 | 0.018181372 |
| Srgn | 0.009132516 | 0.016262546 |
| Stk32c | 0.004145937 | 0.002781108 |
| Stk35 | 0.000918235 | 0.030470246 |
| Stmn1 | -0.09903711 | 0.014209625 |
| Stmn2 | -0.06136074 | -0.01365971 |
| Stox1 | 0.004526238 | 0.022984631 |
| Stra8 | -0.014611494 | 0.019948548 |
| Stx3 | 0.048215445 | -0.02196152 |
| Sycp3 | 0.007235867 | 0.006303238 |
| Synm | 0.014311908 | 0.006702289 |
| Taf7l | 0.012891867 | 0.012279737 |
| Tagln | 0.114366874 | -0.027904114 |
| Tagln2 | 0.05336517 | 0.000763425 |
| Tapbpl | -0.009307726 | 0.009217081 |
| Tax1bp3 | 0.06117966 | 0.003295737 |
| Tbc1d12 | 0.016404925 | 0.047058746 |
| Tbc1d24 | -0.006835035 | 0.002130718 |

| | | |
|---|---|---|
| Tbx3 | -0.04139424 | 0.04918426 |
| Tc2n | 0.002749162 | 0.010167855 |
| Tcea3 | -0.06957443 | -0.017842442 |
| Tcf15 | -0.008614832 | -0.051939607 |
| Tchh | 0.011763787 | 0.01613659 |
| Tcl1 | -0.053692754 | 0.00692011 |
| Tctn1 | -0.000685755 | 0.005643691 |
| Tdgf1 | -0.03347254 | -0.030838102 |
| Tdh | -0.05223236 | -0.007554072 |
| Terf2ip | 0.000758775 | 0.002160812 |
| Tet1 | -0.061818626 | -0.015976204 |
| Tex101 | -0.000150945 | 0.011464384 |
| Tex19.1 | -0.011107044 | 0.040944066 |
| Tfap2c | 0.006706105 | -0.00310314 |
| Tfpi | -0.008907857 | 0.003504205 |
| Tgfb1i1 | 0.03312508 | 0.009040573 |
| Thbs1 | 0.050707042 | -0.003500769 |
| Thnsl2 | 0.006680115 | 0.04814405 |
| Ticrr | -0.019791996 | 0.036941253 |
| Timp2 | 0.033865947 | 0.007101935 |
| Tinagl1 | 0.062086485 | 0.001663974 |
| Tmcc1 | 0.006608349 | 0.002687044 |
| Tmem191c | -0.010451432 | 0.036227405 |
| Tmem245 | -0.03922526 | 0.011106262 |
| Tmem37 | 0.018643016 | -0.002044049 |
| Tmem40 | -0.021743324 | 0.002066627 |
| Tmem59 | 0.03584268 | 0.019605484 |
| Tmem92 | 0.017165937 | 0.10026781 |
| Tmsb4x | -0.013534682 | -0.012368309 |
| Tnfaip6 | 0.00617615 | 0.004526826 |
| Tnfrsf12a | 0.06504678 | -0.007175474 |
| Top2a | -0.01578211 | -0.007125981 |
| Tpbg | 0.038690757 | 0.001831204 |
| Tpi1 | 0.017254356 | -0.043183923 |
| Tpm1 | 0.09352723 | -0.034005668 |
| Tpm4 | 0.07803035 | -0.025431545 |
| Tppp3 | -0.002905848 | 0.013985018 |
| Trap1a | -0.026771378 | 0.004652328 |

| | | |
|---|---|---|
| Trh | -0.018443774 | -0.024344252 |
| Trib3 | 0.005858367 | -0.006114454 |
| Trim47 | 0.010086296 | -6.39E-05 |
| Trp53i11 | 0.006203337 | -0.007045247 |
| Trp53inp1 | 0.01513976 | 0.025013411 |
| Tsc22d1 | 0.007784756 | 0.010558625 |
| Tsen2 | 0.007208667 | 0.045730494 |
| Ttc39c | 5.46E-05 | 0.008842164 |
| Ttn | -0.017568707 | -0.00347148 |
| Tuba1a | 0.05979431 | -0.012805346 |
| Tuba3a | -0.024786158 | 0.024370926 |
| Tubb2a | 0.003215023 | 0.015644817 |
| Tubb3 | -0.02879702 | 0.011429649 |
| Tubb4b | -0.002364613 | 0.01866532 |
| Tubb5 | 0.0168755 | -0.011440624 |
| Tubb6 | 0.085950784 | -0.011294615 |
| Txnip | 0.010349965 | -0.007884254 |
| Uap1l1 | 0.020307692 | 0.013724774 |
| Ubald2 | -0.031101799 | -0.011048649 |
| Ube2c | -0.02482298 | 0.013243949 |
| Ung | -0.021063453 | -0.014089251 |
| Upp1 | -0.068550006 | 0.008458159 |
| Uqcrb | -0.00653793 | 0.006611616 |
| Usp13 | 0.004012968 | 0.002593527 |
| Usp17la | 0.018881164 | 0.103316836 |
| Usp17lc | 0.013823521 | 0.080655046 |
| Usp19 | -0.018596541 | 0.011165514 |
| Usp26 | 0.006734184 | 0.030654024 |
| Utf1 | -0.035537742 | -0.07570607 |
| Vamp8 | 0.045024212 | 0.002784512 |
| Vgll3 | 0.036183413 | 0.000874722 |
| Vim | -0.015108893 | 0.004227261 |
| Wbp5 | 0.04105525 | 0.002546886 |
| Wfdc2 | 0.020020818 | 0.008969964 |
| Wls | 0.032121085 | 0.022271737 |
| Wnt3 | 0.03449491 | -0.001407496 |
| Wnt4 | 0.04272397 | -0.002134218 |
| Wnt6 | 0.003753461 | 0.005994293 |

| | | |
|---|---|---|
| Wnt7b | 0.030495085 | 4.12E-05 |
| Wtap | -0.010470494 | -0.003760451 |
| Xist | 0.007166265 | 0.008428525 |
| Xlr3a | -0.001680175 | 0.004776655 |
| Ywhaq | -0.001418787 | -0.006331375 |
| Zbtb10 | 0.019264381 | 0.01454067 |
| Zc3hav1 | 0.003201806 | 0.02393366 |
| Zfos1 | 0.008287739 | -0.004308424 |
| Zfp287 | -0.000165213 | -1.47E-05 |
| Zfp296 | -0.024954263 | 0.036445774 |
| Zfp316 | -0.002679332 | 0.004225702 |
| Zfp326 | 0.01877454 | -0.013425171 |
| Zfp42 | -0.1324745 | 0.005638579 |
| Zfp428 | -0.034321375 | -0.028616965 |
| Zfp516 | 0.004431526 | 0.068341635 |
| Zfp560 | 0.018428067 | 0.08156205 |
| Zfp57 | -0.05673658 | 0.002906088 |
| Zfp706 | 0.014492439 | -0.03569876 |
| Zfp809 | 0.009048024 | 0.06818436 |
| Zic2 | -0.001189461 | -0.008576863 |
| Zic5 | 0.00169389 | -0.016164217 |
| Zscan4b | 0.010227794 | 0.06323481 |
| Zscan4c | 0.02656372 | 0.15745132 |
| Zscan4d | 0.024336398 | 0.14619476 |
| Zscan4f | 0.026695814 | 0.15859298 |
| Zyx | 0.080330506 | -0.025026485 |

# Appendix F Gene ontology analysis for PC1 and PC2

Gene Ontology (GO) terms enriched for the top 50 gene loadings for PC1, using 965 highly variable genes as a background:

| **Process:** | | | | | |
|---|---|---|---|---|---|
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0031032 | actomyosin structure organization | 2.97E-07 | 2.20E-03 | 11.47 (901,11,50,7) | Cnn1 - calponin 1 |
| | | | | | Cnn2 - calponin 2 |
| | | | | | Csrp1 - cysteine and glycine-rich protein 1 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Zyx - zyxin |
| | | | | | Krt19 - keratin 19 |
| GO:0030029 | actin filament-based process | 9.14E-05 | 3.38E-01 | 4.38 (901,37,50,9) | Cnn1 - calponin 1 |
| | | | | | Cnn2 - calponin 2 |
| | | | | | Pdlim7 - pdz and lim domain 7 |
| | | | | | Csrp1 - cysteine and glycine-rich protein 1 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Krt19 - keratin 19 |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| | | | | | Zyx - zyxin |
| GO:0045214 | sarcomere organization | 1.16E-04 | 2.87E-01 | 12.01 (901,6,50,4) | Csrp1 - cysteine and glycine-rich protein 1 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Krt19 - keratin 19 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0030036 | actin cytoskeleton organization | 1.17E-04 | 2.17E-01 | 4.81 (901,30,50,8) | Cnn1 - calponin 1 |
| | | | | | Cnn2 - calponin 2 |
| | | | | | Pdlim7 - pdz and lim domain 7 |
| | | | | | Csrp1 - cysteine and glycine-rich protein 1 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Krt19 - keratin 19 |
| | | | | | Zyx - zyxin |
| GO:0007155 | cell adhesion | 9.01E-04 | 1.00E+00 | 3.31 (901,49,50,9) | Anxa3 - annexin a3 |
| | | | | | Lgals1 - lectin, galactose binding, soluble 1 |
| | | | | | Cyr61 - cysteine rich protein 61 |
| | | | | | Tinagl1 - tubulointerstitial nephritis antigen-like 1 |
| | | | | | Ctgf - connective tissue growth factor |
| | | | | | Perp - perp, tp53 apoptosis effector |
| | | | | | Zyx - zyxin |
| | | | | | Tnfrsf12a - tumor necrosis factor receptor superfamily, member 12a |
| | | | | | Prtg - protogenin homolog (gallus gallus) |
| **Function:** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0008092 | cytoskeletal protein binding | 5.38E-04 | 7.41E-01 | 3.00 (901,66,50,11) | |
| | | | | | Arl4c - adp-ribosylation factor-like 4c |
| | | | | | Cnn1 - calponin 1 |
| | | | | | Cnn2 - calponin 2 |
| | | | | | Tpm4 - tropomyosin 4 |
| | | | | | Csrp1 - cysteine and glycine-rich protein 1 |
| | | | | | Pdlim7 - pdz and lim domain 7 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | S100a6 - s100 calcium binding protein a6 (calcyclin) |
| | | | | | Cald1 - caldesmon 1 |
| | | | | | Tagln - transgelin |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| | | | | | Anxa2 - annexin a2 |
| **Component:** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0016327 | apicolateral plasma membrane | 1.16E-04 | 9.96E-02 | 12.01 (901,6,50,4) | Cldn7 - claudin 7 |
| | | | | | Cldn6 - claudin 6 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Krt19 - keratin 19 |
| GO:0099513 | polymeric cytoskeletal fiber | 2.11E-04 | 9.03E-02 | 3.60 (901,50,50,10) | Tubb6 - tubulin, beta 6 class v |
| | | | | | Tpm4 - tropomyosin 4 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Tuba1a - tubulin, alpha 1a |
| | | | | | Nes - nestin |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Krt19 - keratin 19 |
| | | | | | Cald1 - caldesmon 1 |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| | | | | | Krt18 - keratin 18 |
| GO:0005856 | cytoskeleton | 2.17E-04 | 6.20E-02 | 2.74 (901,92,50,14) | Rnf128 - ring finger protein 128 |
| | | | | | Tpm4 - tropomyosin 4 |
| | | | | | Tax1bp3 - tax1 (human t cell leukemia virus type i) binding protein 3 |
| | | | | | Csrp1 - cysteine and glycine-rich protein 1 |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| | | | | | Krt19 - keratin 19 |
| | | | | | Cnn1 - calponin 1 |
| | | | | | Tubb6 - tubulin, beta 6 class v |
| | | | | | Cnn2 - calponin 2 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Pdlim7 - pdz and lim domain 7 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Tuba1a - tubulin, alpha 1a |
| | | | | | Zyx - zyxin |
| GO:0005884 | actin filament | 5.00E-04 | 1.07E-01 | 9.01 (901,8,50,4) | Tpm4 - tropomyosin 4 |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| | | | | | Cald1 - caldesmon 1 |
| GO:0099081 | supramolecular polymer | 6.58E-04 | 1.12E-01 | 3.16 (901,57,50,10) | Tubb6 - tubulin, beta 6 class v |
| | | | | | Tpm4 - tropomyosin 4 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Tuba1a - tubulin, alpha 1a |
| | | | | | Nes - nestin |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Krt19 - keratin 19 |
| | | | | | Cald1 - caldesmon 1 |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| | | | | | Krt18 - keratin 18 |
| GO:0099080 | supramolecular complex | 6.58E-04 | 9.37E-02 | 3.16 (901,57,50,10) | Tubb6 - tubulin, beta 6 class v |
| | | | | | Tpm4 - tropomyosin 4 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Tuba1a - tubulin, alpha 1a |
| | | | | | Nes - nestin |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Krt19 - keratin 19 |
| | | | | | Cald1 - caldesmon 1 |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| | | | | | Krt18 - keratin 18 |
| GO:0099512 | supramolecular fiber | 6.58E-04 | 8.03E-02 | 3.16 (901,57,50,10) | Tubb6 - tubulin, beta 6 class v |
| | | | | | Tpm4 - tropomyosin 4 |
| | | | | | Krt8 - keratin 8 |
| | | | | | Tuba1a - tubulin, alpha 1a |
| | | | | | Nes - nestin |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Krt19 - keratin 19 |
| | | | | | Cald1 - caldesmon 1 |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| | | | | | Krt18 - keratin 18 |
| GO:0015629 | actin cytoskeleton | 8.15E-04 | 8.71E-02 | 4.91 (901,22,50,6) | Tpm4 - tropomyosin 4 |
| | | | | | Tax1bp3 - tax1 (human t cell leukemia virus type i) binding protein 3 |
| | | | | | Pdlim7 - pdz and lim domain 7 |
| | | | | | Csrp1 - cysteine and glycine-rich protein 1 |
| | | | | | Actg1 - actin, gamma, cytoplasmic 1 |
| | | | | | Tpm1 - tropomyosin 1, alpha |
| GO:0001726 | ruffle | 8.63E-04 | 8.20E-02 | 8.01 (901,9,50,4) | Pdlim7 - pdz and lim domain 7 |
| | | | | | S100a6 - s100 calcium binding protein a6 (calcyclin) |
| | | | | | Anxa2 - annexin a2 |
| | | | | | Tnfrsf12a - tumor necrosis factor receptor superfamily, member 12a |
| GO:0031012 | extracellular matrix | 9.01E-04 | 7.70E-02 | 3.31 (901,49,50,9) | Anxa3 - annexin a3 |
| | | | | | Lgals1 - lectin, galactose binding, soluble 1 |
| | | | | | Serpinh1 - serine (or cysteine) peptidase inhibitor, clade h, member 1 |
| | | | | | Cyr61 - cysteine rich protein 61 |
| | | | | | Tinagl1 - tubulointerstitial nephritis antigen-like 1 |
| | | | | | Ctgf - connective tissue growth factor |
| | | | | | Ctsl - cathepsin l |
| | | | | | S100a6 - s100 calcium binding protein a6 (calcyclin) |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Anxa2 - annexin a2 |

Gene Ontology (GO) terms enriched for the top 50 gene loadings for PC2, using 965 highly variable genes as a background:

| **Process:** | None | | | | |
|---|---|---|---|---|---|
| **Function:** | None | | | | |
| **Component:** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0000781 | chromosome, telomeric region | 3.24E-04 | 2.77E-01 | 10.27 (901,9,39,4) | |
| | | | | | Zscan4c - zinc finger and scan domain containing 4c |
| | | | | | Ercc4 - excision repair cross-complementing rodent repair deficiency, complementation group 4 |
| | | | | | Zscan4d - zinc finger and scan domain containing 4d |
| | | | | | Zscan4f - zinc finger and scan domain containing 4f |

# Appendix G MOFA loading values

The following table contains loading values for MOFA factors 1-5 for 965 highly-variable genes in the CRISPRa scRNA-seq screen dataset.

| Gene name | Factor 1 loading values | Factor 2 loading values | Factor 3 loading values | Factor 4 loading values | Factor 5 loading values |
|---|---|---|---|---|---|
| 0610040J01Rik | -1.64E-05 | -9.84E-05 | -0.016708871 | 5.41E-04 | -6.60E-05 |
| 1190005I06Rik | 0.05119792 | -0.055497302 | 0.02832786 | 0.006066923 | 0.018664086 |
| 1500009L16Rik | -0.099843855 | 0.026810357 | 0.009992215 | 0.016430227 | -0.08113623 |
| 1700007K13Rik | -0.102220748 | 0.069969535 | 0.006915592 | -0.019718968 | -0.075686096 |
| 1700029P11Rik | -0.014803427 | 0.012537436 | -0.018402045 | -0.036534802 | -0.0131944 |
| 1700052K11Rik | -0.001807191 | 3.20E-04 | -0.022247211 | 0.005591312 | -0.001746657 |
| 1700097N02Rik | -0.033606618 | -0.041147103 | 0.014371641 | -0.081424078 | -0.063518977 |
| 1810032O08Rik | -0.00205713 | -0.007570531 | 0.004912318 | -0.00331627 | -0.003940054 |
| 2200002D01Rik | -9.41E-04 | -0.014305994 | 0.006258242 | 0.086622394 | -0.010764645 |
| 2310043M15Rik | -0.002837845 | -6.94E-05 | 6.30E-04 | -0.003622672 | -0.002307665 |
| 2810417H13Rik | 0.069778357 | -0.061823939 | 0.061534978 | 0.077887059 | -0.011982945 |
| 3830417A13Rik | -7.38E-05 | 4.93E-05 | -7.24E-04 | 0.001439154 | -5.20E-05 |
| 4930548H24Rik | 2.18E-04 | -0.001324998 | -0.002660038 | 3.12E-04 | -4.78E-04 |
| 4933402E13Rik | -2.59E-04 | 1.90E-04 | -0.001327832 | 0.001698329 | -1.17E-04 |
| 5430402O13Rik | -5.26E-05 | 3.65E-04 | -0.048736871 | 7.68E-05 | 1.18E-04 |
| 5430416N02Rik | -0.001927967 | -0.006058275 | 0.002302399 | 0.002743464 | -0.003502204 |
| 6330410L21Rik | 0.001678352 | 0.003991244 | -0.002655588 | -0.012551923 | 0.007723574 |

| | | | | | |
|---|---|---|---|---|---|
| 9530053A07Rik | -0.02472026 | 0.01553811 | 0.001332726 | 0.001406728 | -0.018665551 |
| 9530059O14Rik | -5.70E-04 | 3.34E-04 | -0.053496991 | 1.85E-04 | -6.43E-05 |
| A330069E16Rik | -0.001997408 | -0.018511388 | -0.001286207 | -0.001319213 | -0.006671843 |
| A930014E10Rik | -5.34E-05 | 4.75E-05 | -0.00302827 | 5.03E-04 | -1.65E-05 |
| AA467197 | -0.00361004 | 0.003004582 | -0.003688625 | 0.001503609 | -0.001034348 |
| Aard | -0.023992817 | 0.013084707 | -0.024272381 | -0.167875425 | -0.042255623 |
| Aass | 0.008013066 | -0.006829309 | 0.004559668 | -5.41E-04 | 0.008306242 |
| Abracl | 0.030038778 | -0.032464693 | -0.004119042 | 0.093742596 | 0.004510957 |
| Acaa2 | 0.006275876 | 0.01464349 | 0.004809786 | -0.024980532 | 0.009881799 |
| Acot1 | -0.015956037 | 0.020631897 | -0.016271576 | 0.014945556 | -0.009259785 |
| Acta1 | -2.01E-05 | -4.55E-05 | -6.72E-04 | 0.001052389 | -9.07E-05 |
| Acta2 | 4.12E-04 | -6.73E-04 | -8.33E-04 | 0.006928899 | -1.35E-04 |
| Actb | 0.185524642 | -0.260021805 | 0.027348548 | 0.160365143 | -0.234287627 |
| Actg1 | 0.163897598 | 0.006315632 | 0.022374611 | 0.200243551 | 0.130849949 |
| Acy3 | -9.10E-04 | 0.001850555 | -0.027132225 | 8.10E-05 | -6.28E-04 |
| Adamts9 | -7.18E-05 | -3.68E-04 | -0.001208644 | 0.001404879 | -1.65E-04 |
| Adcy2 | -1.83E-04 | 3.50E-04 | -0.036407559 | -2.14E-05 | 1.47E-05 |
| Adgrl2 | 0.001094054 | -0.016443337 | 0.003763378 | 0.043593273 | -0.00287388 |
| Adgrl3 | -9.43E-05 | 1.79E-05 | -0.001792997 | 5.41E-04 | -7.59E-05 |
| Adprh | -0.005663084 | 0.014169666 | 0.047056789 | -0.134063551 | -0.012227003 |
| Adrb2 | -0.012864867 | 0.009640224 | -6.88E-04 | 0.003452909 | -0.008829507 |
| Aebp2 | -0.01654982 | -0.001489818 | 0.017720717 | -0.044559306 | -0.007083431 |
| Aes | 0.00118272 | -0.025214079 | 0.027818741 | -0.19115023 | -0.027114676 |
| AF067061 | 2.96E-04 | -3.01E-04 | -0.042983776 | -6.62E-05 | 5.40E-05 |
| Agl | 0.017564884 | -0.062610906 | -0.009114699 | -0.052826497 | -0.010837642 |
| Agpat4 | -0.048705255 | 0.018957604 | 0.008871251 | 0.055619647 | -0.035722264 |
| Agpat9 | -1.29E-04 | 2.47E-04 | -0.001014957 | 0.002548952 | 4.76E-06 |
| Ahnak | -0.009616902 | -0.04121678 | 0.008523866 | 0.06846859 | -0.029882448 |
| Ahnak2 | -0.032093106 | 0.012154493 | 0.003579767 | 0.022093877 | -0.01930174 |
| Aire | 0.01877474 | -0.030851763 | 0.024518923 | -0.07970355 | -2.18E-04 |
| Ak4 | 1.07E-04 | -4.82E-04 | -0.01489215 | -0.026441386 | 0.0024795 |
| Akap12 | -0.001492418 | -0.113627564 | 0.031985218 | 0.055437059 | -0.004030306 |
| Akr1b3 | 0.026679975 | 0.008677818 | 0.01020011 | -0.018599475 | -6.57E-04 |
| Aldoa | 0.106355747 | 0.099319229 | 0.011973151 | 0.060156096 | 0.132963707 |
| Amn | 0.001128016 | 9.61E-04 | -0.001400202 | 0.002987181 | 8.93E-04 |

| | | | | | |
|---|---|---|---|---|---|
| Amotl1 | -0.010272671 | -0.002236653 | 0.002878633 | 0.036687233 | -0.00867589 |
| Amotl2 | 0.004310303 | -0.018334789 | -0.005197745 | 0.054587161 | -0.003253642 |
| Amt | 0.037337771 | -0.08022677 | 0.004366164 | -0.001216859 | -0.095884861 |
| Anp32a | 0.022172238 | 0.035746192 | 0.053190988 | -0.042024858 | 0.045445557 |
| Anxa1 | -0.001825195 | 5.41E-05 | -1.59E-04 | 0.021298232 | -0.002816759 |
| Anxa2 | -0.004877658 | 0.01566929 | -0.01351174 | 0.091388141 | -7.15E-04 |
| Anxa3 | -0.002244607 | -0.005473353 | -0.002374091 | 0.132250185 | -0.008042745 |
| Anxa5 | -0.00395981 | 0.003830847 | -0.002670417 | 0.019371857 | -0.00353341 |
| Anxa8 | -6.96E-04 | 4.83E-04 | -8.55E-04 | 8.20E-04 | -3.59E-04 |
| Ap3b2 | -1.91E-04 | 8.69E-05 | -0.001247699 | 0.001485579 | -2.37E-05 |
| Apela | 0.011321697 | -0.046333351 | 0.011182686 | 0.109734898 | -0.008428776 |
| Apoa2 | -7.21E-05 | 5.03E-05 | -0.00203238 | 6.38E-04 | -5.80E-05 |
| Apobec2 | 7.69E-04 | 3.61E-04 | -0.004198215 | 0.005607919 | 6.72E-04 |
| Apoe | -0.038010405 | 0.094175554 | -0.0130344 | -0.012431727 | 0.041556955 |
| Aqp3 | -0.004637531 | 0.025780196 | -0.006612321 | -0.01915915 | 0.012297574 |
| Arg2 | 4.99E-05 | 0.00291273 | -0.115706891 | 0.001112312 | 4.03E-04 |
| Arhgef26 | 3.80E-04 | -0.002239583 | -0.076060279 | 0.00394275 | -9.57E-04 |
| Arl14epl | 0.014514634 | 0.008548342 | 0.007901834 | -0.061056098 | 0.017077692 |
| Arl4a | -0.035872277 | -0.012091445 | 0.008096638 | -0.018563717 | -0.026676897 |
| Arl4c | 0.002471111 | -0.017525015 | 0.005553532 | 0.097091046 | -0.004919775 |
| Arl4d | 0.004362126 | 0.00106709 | -0.0115497 | -0.00748112 | 0.003701525 |
| Arl6ip1 | -0.035846006 | -0.002192206 | -0.027969643 | -0.03481223 | 0.074135384 |
| Arpc1b | 0.050637016 | -0.003534648 | 0.020058528 | 0.053152135 | 0.042730297 |
| Ascl2 | 0.001031246 | 8.29E-04 | -0.003153685 | 0.015068226 | -1.66E-04 |
| Asns | 0.032128593 | 0.120271989 | -0.038636024 | 0.014616779 | 0.087172986 |
| Ass1 | -0.06141201 | 0.096419087 | -0.004524482 | -0.035667525 | -0.017303248 |
| Atf4 | 0.002802434 | -0.050677254 | -0.00170537 | 0.015387163 | 0.013235908 |
| Atg2a | -0.003283844 | -0.011595056 | -1.17E-04 | -0.016090115 | -0.005514998 |
| Atp5b | -0.001555455 | 0.169564514 | -0.028148154 | 0.001730636 | 0.084253072 |
| Atp6v0b | -0.058097083 | 0.014002742 | 0.00258549 | 0.053292396 | -0.03496963 |
| Atp6v0d1 | -0.03864474 | 0.04750481 | -0.015944998 | 0.003327844 | -0.01620655 |
| Atrn | -0.002468902 | -0.004014658 | -0.010249849 | -0.001487685 | -0.003281419 |
| Atrx | 0.011781314 | -0.083626 | 0.028975241 | -0.004295632 | 0.025335327 |
| Avpi1 | -0.023189299 | 0.025029272 | -0.127308507 | 0.032261125 | -0.003650807 |
| AY036118 | -0.002329112 | -0.003237887 | -0.001463861 | -0.003107197 | -0.002034387 |
| B020031M17Rik | 5.92E-04 | -7.51E-04 | -0.094365551 | -4.45E-04 | 1.96E-04 |
| B2m | -0.012814985 | 0.004730819 | -3.17E-04 | 0.082992882 | -0.011890513 |

| | | | | | |
|---|---|---|---|---|---|
| B930036N10Rik | 7.11E-04 | -0.006083581 | -0.00147169 | 0.001976699 | -9.48E-04 |
| Bambi | -7.40E-04 | -7.57E-04 | -0.020336492 | 0.017280632 | 0.002828387 |
| Baz1a | 0.011985194 | -0.011778463 | -0.003656163 | 0.08009866 | 0.005016219 |
| Bbc3 | -0.086464745 | 0.054541518 | 0.022943039 | 0.001898662 | -0.076642978 |
| Bbs2 | -0.007855784 | -0.024344185 | 0.015533784 | -0.081017872 | -0.039494291 |
| Bbx | -0.007629905 | -0.013349933 | 0.007197867 | 0.105158049 | -0.006844748 |
| BC028528 | -0.002266885 | -0.004569495 | 0.001771613 | -0.013842376 | -0.003620539 |
| BC080695 | 8.47E-04 | -5.15E-05 | -0.090865973 | -1.18E-04 | -5.87E-05 |
| BC147527 | -5.20E-06 | -2.13E-05 | -0.001510012 | 2.51E-04 | -2.36E-05 |
| Bcat1 | 0.066568095 | 0.002567625 | 0.007304005 | -0.064860599 | 0.071445408 |
| Bdnf | -4.62E-05 | -1.86E-04 | -0.003648481 | 0.001951382 | 1.67E-05 |
| Bend7 | -2.56E-04 | 6.99E-05 | -8.89E-04 | 0.001360025 | -1.91E-04 |
| Bex1 | 0.075791679 | 0.015181976 | 0.027972402 | 0.156026497 | 0.057246242 |
| Bhlhe40 | -0.001253216 | -0.001541938 | 5.20E-04 | 0.023965158 | -0.002428279 |
| Bmp4 | -0.001762911 | -0.006057639 | 0.008490173 | -0.025918458 | -0.004612275 |
| Bnip3 | 0.00437431 | 0.042005337 | -0.013319041 | -0.03380037 | 0.022067611 |
| Bst2 | 0.00959395 | -0.03967841 | 0.014500556 | 0.122353748 | -0.019295127 |
| Btg1 | -0.062471779 | 0.038542377 | -0.025140529 | 0.018166291 | -0.039578853 |
| Btg2 | -0.103135899 | 0.065864774 | -0.037684591 | 0.027934585 | -0.090672492 |
| C130073F10Rik | -0.003645691 | -1.92E-04 | -0.008364591 | -0.003736167 | -0.001273319 |
| C1d | 0.029036575 | 0.003053305 | -0.098752382 | 0.029763534 | 0.019838323 |
| C2cd4b | -1.93E-04 | 1.95E-04 | -0.005990939 | 0.001001334 | -9.93E-05 |
| C730034F03Rik | -4.34E-04 | -5.36E-04 | -0.001654719 | 0.001681337 | -4.77E-04 |
| C77370 | 8.72E-05 | -0.006644547 | 0.009253659 | -0.015318378 | 0.005788168 |
| Cacybp | 0.114943413 | -0.032354908 | 0.025161684 | -0.086614299 | 0.091770877 |
| Calb2 | -2.82E-04 | 6.38E-04 | -0.001076572 | -0.002784236 | -1.37E-04 |
| Calcoco2 | 0.026686876 | 0.073174121 | -0.044628411 | -0.105874532 | 0.097703174 |
| Cald1 | -0.064893255 | 0.029559859 | 8.50E-04 | 0.160028779 | -0.045335754 |
| Calm2 | 0.067187566 | 0.036966333 | 0.030826652 | 0.036326672 | 0.09192958 |
| Calml4 | 0.005869401 | 0.001019568 | 0.002941414 | 0.001801605 | 0.0073673 |
| Calr | 0.050558014 | -0.039195262 | 0.004973551 | 0.011280102 | 0.039704829 |
| Camk2d | -0.001626211 | -5.70E-04 | -4.36E-05 | 0.011799522 | -0.001818141 |
| Capg | -0.012250891 | 0.019652057 | 0.00603985 | 0.047409007 | -0.004070248 |
| Car2 | 0.041337663 | -0.086881879 | 0.058758336 | 0.349020991 | 0.008670596 |
| Car3 | 0.004953044 | -0.013737058 | 0.003347546 | 0.0342049 | 0.001489169 |

| | | | | | |
|---|---|---|---|---|---|
| Car4 | 0.010810812 | -0.010699766 | -0.002242319 | 0.064545323 | 0.002595662 |
| Carhsp1 | -0.038166515 | -0.002096224 | 0.022419103 | 0.051148046 | -0.050820425 |
| Catip | -0.006384537 | -9.89E-04 | 0.005187458 | -0.024629173 | -0.006991911 |
| Cbr3 | -0.002731243 | 1.29E-04 | 0.001020291 | 0.01587 | -0.004288248 |
| Cbx3 | 0.18293127 | -0.340606946 | 0.039480465 | 0.049133992 | -0.314584406 |
| Ccnb1 | 0.021258525 | 0.061726301 | -0.022735935 | 0.002239191 | 0.095880172 |
| Ccnd1 | 0.033109442 | -0.051319329 | 0.027216622 | -0.019526893 | 0.011123962 |
| Ccnd2 | 0.023391761 | -0.018852645 | 0.00640529 | -7.33E-04 | 0.018103764 |
| Ccnd3 | 0.075778782 | -0.033940822 | 0.036064228 | -0.109799317 | 0.023242611 |
| Ccne1 | 0.190388557 | -0.078209438 | 0.048434863 | -0.011164901 | 0.059286369 |
| Ccne2 | 0.01798876 | -0.006017423 | -0.00938027 | 0.016004519 | 0.001848098 |
| Ccng1 | -0.166363859 | 0.084038422 | -0.019548519 | 0.059449155 | -0.148642289 |
| Ccno | 0.002777503 | 0.002827105 | -0.031949444 | 0.003350482 | 0.004287386 |
| Ccser1 | 2.38E-04 | 3.66E-04 | -0.080676572 | -3.24E-04 | 2.39E-04 |
| Cd24a | -0.009494851 | 0.003388046 | -0.017272257 | 0.080675032 | -0.009325496 |
| Cd37 | 9.37E-04 | 9.49E-05 | 0.005142041 | -0.0208592 | 3.53E-04 |
| Cd63 | -0.00482081 | 0.057713437 | 0.00812137 | 0.134095529 | 0.044366772 |
| Cd74 | 0.002466542 | -0.001430364 | -0.006320924 | -0.014402953 | -0.003457686 |
| Cd81 | -0.111158738 | 0.139024058 | -0.009013965 | 0.083756046 | -0.064736097 |
| Cd9 | 0.01392375 | 0.098358928 | 0.01634333 | 0.001808993 | 0.077954983 |
| Cdc14a | -4.86E-04 | -3.08E-05 | -0.001601476 | 0.001833864 | -4.30E-04 |
| Cdc20 | 0.087950049 | 0.076920228 | -0.010947202 | 0.031777174 | 0.089531418 |
| Cdc42ep3 | -0.015303487 | 0.009276706 | -0.027703591 | -0.0128784 | -0.004486343 |
| Cdc5l | 0.004396074 | 0.036710642 | -0.024749575 | -0.154174685 | 0.035419267 |
| Cdk6 | 0.014269485 | -0.035806689 | 0.006811119 | 0.045769295 | -0.009760534 |
| Cdkn1a | -0.112470369 | 0.065875039 | -0.002680124 | 0.052702553 | -0.096877639 |
| Cdkn1c | 2.35E-05 | -0.007498157 | 0.003670972 | 0.056659389 | -0.009525403 |
| Cdkn2a | -0.008543306 | 0.010760173 | -0.002264194 | 0.034806426 | -0.004571139 |
| Cdv3 | 0.111625495 | -0.178043138 | 0.002600457 | 0.085420415 | 0.012816608 |
| Cdx2 | 1.38E-05 | -2.89E-04 | -0.001072514 | 0.004627946 | -4.05E-04 |
| Cebpb | -0.005275136 | 0.00752378 | -0.002990156 | 0.054124281 | -0.004970284 |
| Cenpf | -0.057609275 | -0.081664483 | 0.001261754 | -0.03394279 | 0.006619383 |
| Cenpm | -0.046140418 | 0.038331668 | -0.01097005 | -0.148012602 | -0.027369368 |
| Cep57l1 | -0.005872355 | -0.010028269 | -0.054425886 | -0.020997424 | 1.19E-04 |
| Chac1 | 0.006620268 | -0.011173837 | 0.004714957 | 0.013742636 | 0.011203284 |
| Chchd10 | -0.06277249 | 0.001255989 | 0.020071699 | -0.246962329 | -0.029290169 |
| Cidea | -0.001622151 | 0.010059251 | -0.017860186 | -0.005514051 | 0.002995881 |
| Cited1 | 0.001857397 | -0.001907073 | -0.004654795 | 0.022570736 | 4.57E-04 |

| | | | | | |
|---|---|---|---|---|---|
| Cited2 | 0.009763449 | -0.021046297 | -0.014136973 | 0.138729496 | 0.024654853 |
| Cited4 | 7.52E-04 | 0.001866525 | -3.06E-04 | 0.008993038 | 0.001814248 |
| Ckb | 0.075895139 | 0.104520841 | -0.016974494 | -0.009886863 | 0.150356718 |
| Ckmt1 | -0.004301252 | 0.00515485 | -0.00248932 | 0.001449239 | -0.00264495 |
| Clcn5 | -3.06E-04 | -0.001041867 | -0.008382587 | 0.016843902 | 1.81E-04 |
| Cldn3 | -6.88E-04 | 6.12E-04 | 7.55E-04 | 0.010941076 | -5.97E-04 |
| Cldn4 | -0.008349068 | 0.006551779 | 0.006819324 | 0.042639667 | 9.55E-04 |
| Cldn6 | 0.007970531 | 0.004301318 | 4.12E-04 | 0.144475801 | 0.005934536 |
| Cldn7 | 0.002019527 | -0.00559098 | 0.005333633 | 0.075816932 | -4.64E-05 |
| Clic1 | -0.062706873 | 0.049709351 | 0.01187332 | 0.113195122 | -0.049723933 |
| Clic4 | 0.005027284 | -0.020984948 | -0.00682073 | 0.018430925 | -0.008756151 |
| Clp1 | 0.020959219 | 0.024493481 | -0.087804235 | -0.005040113 | 0.02137188 |
| Cltb | -0.025097673 | 0.001557435 | 0.020152524 | -0.082565099 | -0.038199095 |
| Clu | 0.001864753 | 0.003616442 | -0.003694211 | 0.029194464 | 3.31E-04 |
| Cmtm7 | -0.006416326 | -0.003417852 | 0.027168573 | 0.006193871 | -0.014589284 |
| Cnn1 | 0.005812599 | -0.007792661 | 9.67E-04 | 0.047561496 | 8.98E-04 |
| Cnn2 | 0.071298209 | -0.03914387 | 0.020733788 | 0.172104957 | 0.061669196 |
| Cntfr | 0.004547623 | -0.01399491 | 0.004273464 | 0.029474866 | -0.010397838 |
| Coch | 0.004214286 | -0.001758467 | -0.025456005 | 0.007053788 | 0.002796604 |
| Col4a1 | -0.001936047 | -1.31E-04 | -0.002072597 | 0.010337535 | -0.002571526 |
| Col4a2 | -7.84E-04 | -1.86E-04 | -8.41E-04 | 0.008160315 | -0.001226698 |
| Commd3 | -0.061359348 | 0.047290875 | -0.008445992 | 0.024449325 | -0.040372128 |
| Comt | 0.029302007 | -0.117933642 | -0.022506839 | -0.067444799 | -0.039429431 |
| Cox6b2 | -0.115313962 | 0.043469447 | -0.040333891 | -0.063817788 | -0.090911446 |
| Cpsf4l | 0.014417819 | -0.007198232 | 0.00954525 | -0.019527614 | 0.027546795 |
| Crabp2 | 0.011253593 | -0.003325621 | -0.001673242 | 0.022559956 | 0.008002649 |
| Crct1 | 0.002470118 | -0.007623741 | -0.013054993 | 0.00806115 | -0.003679839 |
| Crim1 | -2.42E-04 | -1.71E-04 | -0.001166685 | 0.001755015 | -3.12E-04 |
| Crip2 | -0.051549029 | 0.027201374 | 0.04296773 | 0.109229543 | -0.065715848 |
| Crxos | -4.95E-06 | 0.00547063 | -0.019324182 | -0.00583892 | 0.004120918 |
| Cryab | -0.010496794 | 0.003444534 | 0.001364188 | 0.044319273 | -0.011568574 |
| Csrp1 | -0.003747568 | -0.001617014 | 0.005473499 | 0.121617278 | -0.007689416 |
| Cst3 | 0.020167744 | 0.028125188 | -0.011324862 | 0.084444342 | 0.044340143 |
| Cstb | -0.032388744 | 0.015350976 | -0.004244022 | 0.097094986 | -0.028571709 |
| Ctgf | 0.001137565 | -0.011562482 | -0.013963643 | 0.088890145 | 0.004267792 |
| Cth | 0.023872198 | 0.033273466 | 0.005603243 | 4.32E-04 | 0.032896569 |
| Ctnnal1 | 0.052782033 | -0.015391754 | 0.028128439 | -0.033552404 | 0.058139189 |
| Ctr9 | -0.014259479 | -0.005184157 | -0.069696067 | 0.033986501 | -0.005137919 |

| | | | | | |
|---|---|---|---|---|---|
| Ctsa | -0.01442814 | 0.019019711 | 0.003007513 | 0.00764014 | 0.001429553 |
| Ctsb | -0.010068228 | 0.016334212 | -0.005566725 | 0.016693219 | -0.002812492 |
| Ctsd | -0.080968805 | 0.084760497 | 0.019164551 | 0.02515032 | -0.039431559 |
| Ctsl | -3.71E-04 | 0.034242172 | -0.024620947 | 0.145818651 | 0.025270595 |
| Cul9 | -2.54E-04 | -0.001044932 | -0.001044119 | -0.001007928 | -8.38E-04 |
| Cwc22 | 0.004889109 | -0.017037284 | -0.098902106 | -0.001607839 | 0.002220599 |
| Cxcl16 | -0.003050792 | -9.02E-04 | 5.35E-04 | 0.026800654 | -0.003913324 |
| Cyba | -0.013571377 | 0.011991322 | 0.005310678 | 0.06817394 | -0.009640359 |
| Cyp26a1 | 8.44E-05 | -7.24E-05 | -0.0010308 | 0.001387127 | 5.83E-05 |
| Cyr61 | 0.001946902 | -0.007367544 | 0.001088503 | 0.119287155 | 7.01E-04 |
| D630023 F18Rik | -0.088496611 | 0.047470427 | 0.009283538 | -0.02754632 | -0.06998404 |
| Dab2 | -1.10E-04 | -2.90E-04 | -0.001589461 | 0.008254913 | -4.99E-04 |
| Dazl | -0.001749509 | 0.00706176 | -0.025556418 | -0.002609924 | 0.002464996 |
| Dcaf12l1 | 0.009226574 | -0.009975048 | -0.017155671 | 0.025060714 | 0.004435548 |
| Dcaf17 | 0.007339095 | -0.016986354 | -0.004880772 | 0.006777488 | 0.007277458 |
| Dcbld1 | -0.009354884 | -0.00449964 | -0.040195407 | -0.029576078 | -0.012212218 |
| Dcxr | -0.08874999 | 0.062595346 | -0.003288986 | 0.016213483 | -0.057298704 |
| Ddah1 | 0.022431776 | -0.024526665 | 0.017414382 | 0.085340267 | 0.015701634 |
| Ddah2 | -0.02537903 | 0.039523182 | 5.63E-04 | -0.014811586 | 5.32E-04 |
| Ddit3 | -0.019889465 | 0.011401558 | -0.002285091 | 0.020444058 | -0.008525693 |
| Ddit4 | -0.063952156 | 0.059287985 | -0.04970163 | 0.044531961 | -0.047330149 |
| Ddit4l | -0.084868929 | 0.058088817 | -0.071576953 | 0.014222924 | -0.068988054 |
| Ddr2 | -9.58E-05 | 6.56E-05 | -0.001571316 | 4.73E-04 | -8.78E-06 |
| Derl3 | -0.012759253 | 0.008609676 | -0.001860873 | -0.006164773 | -0.005419054 |
| Dhx16 | -0.141855551 | 0.1040554 | 0.05850894 | -0.151058479 | -0.124998154 |
| Dkkl1 | -0.001673051 | 8.38E-04 | -1.28E-04 | 0.002103873 | -0.001990243 |
| Dlg1 | 0.002677652 | -0.026914269 | -0.008514434 | -0.007876062 | -0.00693695 |
| Dll1 | -0.001372554 | 4.57E-04 | -0.002911209 | 0.001330751 | -6.97E-04 |
| Dmkn | -4.04E-04 | -2.04E-04 | -0.002437195 | 0.006417728 | -0.001916393 |
| Dnajb14 | -7.45E-04 | -0.004442401 | -0.031119085 | 0.001437798 | -2.28E-04 |
| Dnajb9 | -0.00131515 | 0.00569786 | -0.047458063 | 0.002485221 | 0.001164692 |
| Dnajc21 | 0.050401571 | -0.08743364 | 0.051285653 | -0.123517993 | 0.053203139 |
| Dnmt3a | 0.030918753 | -0.046534992 | 0.01293488 | 0.010919387 | 0.008873098 |
| Dnmt3b | 0.11562517 | -0.137858236 | 0.042192718 | 0.147109069 | 0.064100381 |
| Dnmt3l | 0.068847418 | -0.032045895 | 0.035379875 | -0.036493584 | 0.050372082 |
| Dok2 | 0.009431951 | -0.014051977 | 0.00410573 | 0.038686403 | 0.001822278 |
| Dpp7 | -0.013425823 | 0.007817644 | -3.95E-04 | -7.95E-04 | -0.008026651 |
| Dppa2 | 0.057516227 | 0.042229329 | -0.11396688 | -0.08832436 | 0.050366908 |

| | | | | | |
|---|---|---|---|---|---|
| Dppa3 | 0.026095039 | 0.03368135 | -0.013929964 | 0.020936753 | 0.081884532 |
| Dppa5a | -0.02893406 | 0.09518413 | -0.016501663 | -0.21672944 | 0.050783953 |
| Dpys | -0.024712134 | 0.015927109 | 0.017975594 | -0.040595264 | -0.006596137 |
| Dqx1 | 0.00585105 | -0.025997624 | -0.030776835 | -0.046239427 | -0.00981035 |
| Dsp | -0.002367313 | -0.002714334 | 2.78E-04 | 0.024994979 | -0.003560883 |
| Dst | -0.02388672 | -0.03706837 | -0.072432238 | 0.022951106 | -0.007662257 |
| Dstn | 0.013090822 | -0.008579842 | 0.032338864 | 0.1266342 | -0.005727614 |
| Dusp1 | -0.004468022 | 0.004969258 | -0.009131405 | 0.050361787 | -0.005110158 |
| Dusp14 | 0.003024268 | -0.001512781 | 0.001650585 | 0.027309719 | -0.001048101 |
| Dusp4 | 0.00292833 | -0.00873667 | -0.001365125 | 0.05752108 | -0.003170862 |
| Dusp5 | 0.002438329 | -0.003871727 | -0.009834512 | 0.028974363 | -0.002220012 |
| Dusp6 | 0.0105146 | -0.01819276 | 0.00832839 | 0.076903275 | -0.00736267 |
| Dusp9 | 0.00573718 | -0.010671937 | -0.023372183 | 0.01263387 | -0.016531377 |
| Duxf3 | 1.59E-04 | 9.56E-05 | -0.021762989 | 3.76E-04 | -7.19E-05 |
| Dynll1 | 0.064217295 | -0.022591575 | -0.012059276 | -0.011963725 | 0.021296379 |
| E130012 A19Rik | 0.051030588 | -0.053908744 | 0.066409148 | -0.275357287 | -0.008035156 |
| E2f5 | 0.006537197 | -0.008935749 | 0.002289885 | 0.017406748 | 0.002717996 |
| Ebf1 | -0.004717407 | -0.003280733 | 0.00109076 | 0.004869667 | -0.005410788 |
| Echdc2 | 0.011467733 | 0.021421721 | -0.022897931 | -0.009221447 | 0.020690461 |
| Echdc3 | -4.52E-05 | -1.13E-05 | -0.021986734 | 3.74E-04 | -3.33E-05 |
| Edn1 | -4.74E-04 | -2.09E-04 | -0.001223535 | 0.003380792 | -1.75E-04 |
| Edn2 | 1.26E-04 | -1.98E-04 | -9.77E-04 | 5.18E-04 | -1.52E-04 |
| Eef1a1 | -0.140210412 | 0.175179446 | -0.049251631 | 0.010052744 | -0.059237215 |
| Eef2 | -0.100409133 | 0.103721978 | -0.01714944 | 0.0518017 | -0.043488528 |
| Efhc2 | -4.16E-05 | -8.41E-07 | -0.001460849 | 0.001045154 | 1.24E-04 |
| Efna3 | -0.008828266 | 0.009044213 | -0.001817414 | 0.00537098 | -0.00865042 |
| Efnb1 | -2.12E-04 | -9.97E-05 | -0.001100262 | 0.003539295 | -4.15E-04 |
| Egr1 | -4.35E-04 | -0.007907884 | 0.003715727 | -0.001388342 | -0.007010252 |
| Egr4 | 3.95E-05 | -8.56E-05 | -7.60E-04 | 7.86E-04 | -8.32E-05 |
| Eif2s2 | 0.038484869 | 0.036082226 | -0.01469929 | -0.166231104 | 0.075827411 |
| Eif4a1 | 0.106236302 | 0.126207131 | -0.01019854 | 0.032788772 | 0.138068741 |
| Elf3 | 0.00194371 | 0.011899388 | -0.001938691 | 0.011812553 | 0.008117363 |
| Emb | 0.108335994 | -0.059511556 | 0.016072802 | 0.079885442 | 0.082800503 |
| Emilin1 | 0.003916392 | -0.023667678 | -0.005229164 | -0.026148734 | -0.005798606 |
| Emp2 | -3.24E-04 | 1.33E-04 | -7.09E-04 | 0.002187796 | -3.15E-04 |
| Emp3 | -0.013887099 | 0.010254064 | 9.14E-04 | 0.024656331 | -0.014658998 |
| En1 | 9.80E-04 | -0.00444255 | 8.87E-04 | 6.58E-04 | 1.09E-04 |
| Enah | 0.0113391 | -0.128684911 | 0.034236129 | -0.146979717 | -0.00802233 |

| | | | | | |
|---|---|---|---|---|---|
| Endov | 2.73E-04 | -0.004573663 | 8.76E-04 | 0.001001325 | -6.66E-04 |
| Eno1 | 0.090565349 | 0.111180958 | 0.026734429 | 0.100715697 | 0.149717013 |
| Enpp2 | 1.86E-04 | -7.66E-04 | -0.001211657 | 0.003475875 | 1.26E-04 |
| Epcam | 0.034010681 | -0.029188862 | 0.021021746 | 0.163349484 | 0.027477528 |
| Epha2 | -0.020066724 | -0.028438934 | 0.01496961 | 0.04716326 | -0.032603934 |
| Eps8l2 | -0.004114629 | 0.004014838 | -1.42E-04 | 0.012905676 | -0.001592546 |
| Ercc4 | -0.002215061 | 0.001245847 | -0.095417966 | -0.003868583 | 6.64E-04 |
| Erdr1 | -0.020302122 | -0.057958694 | 0.026487699 | 0.092811499 | -0.011242796 |
| Esco2 | -0.03554489 | -2.99E-04 | -0.010796055 | -0.118914886 | -0.04200793 |
| Esx1 | -1.14E-04 | 1.77E-04 | -0.001358406 | 0.002146922 | -9.67E-05 |
| Eva1a | -1.80E-05 | -1.82E-05 | -0.00209676 | 8.64E-04 | -8.12E-05 |
| F2r | 0.00330905 | -0.012413867 | 0.0038407 | 0.044122063 | -0.003526789 |
| F2rl1 | -0.001055874 | -0.028717474 | 0.013969573 | 0.033875021 | -0.012502779 |
| F3 | 1.83E-04 | -0.003827281 | 0.001274117 | 0.02455504 | -8.13E-04 |
| Fabp3 | -0.032035211 | 0.037614413 | 0.004125562 | 0.029008374 | 0.005081786 |
| Fabp5 | 0.052648691 | -0.001557334 | 0.00738152 | 0.179776791 | 0.055374388 |
| Fam124a | -1.06E-04 | 4.97E-05 | -0.01452543 | 4.40E-04 | -1.14E-04 |
| Fam134b | -3.66E-04 | 2.00E-04 | -0.001117576 | 0.001459293 | -2.19E-04 |
| Fam159b | -1.62E-04 | -8.52E-06 | -8.85E-04 | 0.001029853 | -2.43E-04 |
| Fam162a | -0.070648979 | 0.027871336 | -0.004731923 | -0.03974488 | -0.076636836 |
| Fam25c | -0.035323946 | -0.001665553 | 0.014956403 | -0.130688838 | -0.042410936 |
| Fam60a | 0.072622456 | -0.142330875 | 0.077399796 | 0.059099101 | 0.023296953 |
| Fam83a | -7.59E-05 | -5.97E-06 | -6.35E-04 | 5.14E-04 | -9.43E-05 |
| Fblim1 | 0.035877628 | -0.023352163 | 0.045056506 | -0.092889344 | 0.035888746 |
| Fbp2 | -0.004877602 | 0.006603852 | -0.002595266 | -5.36E-04 | -0.002755656 |
| Fbxo15 | 0.021584545 | 0.0611058 | -0.14410016 | -0.05709737 | 0.081108409 |
| Fgf1 | -7.23E-05 | -1.24E-04 | -0.022319697 | 3.55E-04 | -2.01E-04 |
| Fgf15 | 0.010080361 | -0.005923826 | -0.001176662 | 0.002323693 | 0.006070613 |
| Fgf17 | 0.02450941 | -0.024001795 | 0.00922773 | -0.011281362 | 0.014615533 |
| Fgf4 | 0.001309944 | -0.014353055 | 0.015750288 | -0.080578117 | 0.018723597 |
| Fgf5 | 4.71E-05 | -1.36E-04 | -8.25E-04 | 0.001690417 | -5.84E-05 |
| Fgf8 | -3.90E-05 | -4.66E-05 | -8.58E-04 | 0.001054681 | -9.98E-05 |
| Fgfbp1 | -0.003392568 | -0.011203504 | -0.001590465 | 0.144939248 | 0.00534745 |
| Fhl1 | 0.011882202 | 0.003289324 | 0.004058467 | 0.033192659 | 0.016089584 |
| Fhl2 | -1.27E-04 | -2.02E-04 | -6.87E-04 | 0.003804322 | -3.39E-04 |
| Fkbp6 | 0.001857913 | 0.008625245 | -0.006851525 | -0.016321575 | 0.002091344 |
| Flrt3 | -4.24E-05 | -6.76E-05 | -8.02E-04 | 0.001042444 | -1.09E-04 |
| Fmr1nb | 0.008256103 | 0.014050699 | -0.045457311 | -0.009433069 | 0.015813222 |

| | | | | | |
|---|---|---|---|---|---|
| Folr1 | 0.074007562 | -0.042343129 | 0.023033419 | 0.070394972 | 0.055520708 |
| Fos | -6.34E-04 | 1.77E-04 | -0.002807588 | 0.00357296 | -7.21E-04 |
| Foxo3 | -0.023896826 | -0.01445203 | -0.007575857 | -0.005973554 | -0.075045165 |
| Fst | -0.003209041 | -0.004802358 | 3.86E-04 | 0.023478917 | -0.003992715 |
| Fth1 | 3.59E-04 | -0.095314197 | 0.061827872 | 0.009774133 | -0.054453451 |
| Fthl17a | -1.37E-04 | 1.24E-04 | -0.001318714 | 0.001113629 | -1.09E-04 |
| Fthl17c | 3.96E-04 | 7.44E-04 | -0.021149556 | -0.001108353 | -5.47E-04 |
| Fuca2 | -0.002394649 | 0.001823138 | 4.76E-04 | 0.006601601 | -3.20E-04 |
| Fundc1 | 0.005905445 | 9.56E-04 | -0.065254363 | 0.013147196 | 0.0066247 |
| Fxyd6 | 0.01360516 | 0.019860069 | 0.033838542 | -0.004019895 | 0.053804333 |
| G0s2 | -2.33E-04 | -4.88E-04 | -3.53E-04 | 6.05E-05 | 2.23E-04 |
| Gabarapl2 | 0.004545074 | 0.066243233 | -0.066746752 | -0.316302345 | 0.019574501 |
| Gadd45a | 0.010510289 | -0.011148653 | 0.024239976 | 0.040168359 | 0.020691004 |
| Gadd45b | 0.01522284 | 0.004117832 | -0.009637153 | 0.041455061 | 0.017838244 |
| Gadd45g | -0.014587589 | 0.008225759 | 0.005179508 | 0.024498606 | 0.002171757 |
| Gap43 | 0.003519096 | -0.011485812 | 0.007926244 | 0.037327157 | 6.01E-04 |
| Gata2 | -1.55E-04 | 1.07E-04 | -0.001571539 | 0.002431539 | -1.02E-04 |
| Gata3 | 9.07E-05 | 1.97E-05 | -0.003622361 | 0.010258101 | -3.92E-04 |
| Gbp2 | 3.83E-06 | -2.50E-04 | -9.86E-04 | 0.002791261 | -1.60E-04 |
| Gbp2b | 3.22E-06 | -1.54E-04 | -7.79E-04 | 0.001818922 | -1.01E-04 |
| Gbx2 | -0.001645666 | -0.055518805 | 0.018906978 | 0.002516833 | -0.038771575 |
| Gchfr | -4.54E-04 | -0.003537102 | 0.0024851 | 0.021491952 | -0.00515539 |
| Gfod2 | -0.003029036 | -0.030180064 | -0.004155773 | -0.006291436 | -0.011461168 |
| Gja1 | 0.001170923 | -0.051351185 | 0.014844467 | 0.093402202 | -0.035534876 |
| Gjb3 | 0.021364087 | 0.010811068 | 0.017993475 | -0.001436533 | 0.044418746 |
| Gjb5 | -0.006634731 | 0.013844334 | 0.012363573 | -0.026846886 | 0.002718672 |
| Glipr2 | 0.013073605 | 0.008864126 | 0.017621548 | 0.074427413 | 0.009956854 |
| Glod5 | 0.01922249 | -0.004226632 | 0.006299184 | -0.010670233 | 0.019875125 |
| Glrx | -9.13E-04 | 0.016369989 | -0.053648073 | 0.011795575 | 0.010180022 |
| Glrx2 | -0.017792132 | -0.008548025 | -0.164835589 | 0.019343227 | -0.016696287 |
| Glul | 0.068909576 | -0.008782651 | -0.026729407 | -0.014334736 | 0.052827571 |
| Gm10116 | -0.059970923 | -0.046734451 | 0.018676674 | 0.009280555 | -0.070551574 |
| Gm11238 | -1.31E-05 | -1.76E-05 | -0.001798697 | 3.77E-04 | -2.93E-05 |
| Gm12794 | 5.56E-04 | -4.19E-04 | -0.037266974 | 1.74E-04 | -8.99E-05 |
| Gm13075 | 0.05620451 | -0.143123775 | -0.010480177 | -0.091262131 | -0.008900889 |
| Gm13119 | 5.65E-04 | 1.31E-04 | -0.06828406 | -6.38E-05 | -7.21E-05 |
| Gm13145 | 0.0111479 | -0.026127387 | -0.021956411 | -0.030896794 | 0.003825941 |
| Gm13225 | 0.036458619 | -0.084718361 | -0.021370165 | -0.100279448 | 0.016418349 |

| | | | | | |
|---|---|---|---|---|---|
| Gm15879 | 0.002964355 | -0.008266023 | -0.004502007 | -9.88E-04 | -8.23E-04 |
| Gm15915 | -4.62E-04 | -7.37E-05 | -0.001136571 | -5.91E-05 | -5.96E-04 |
| Gm16233 | -0.002523228 | -0.018004553 | 0.003031246 | 0.055315879 | -0.005475105 |
| Gm2016 | 7.94E-04 | 7.28E-04 | -0.123628856 | -0.001200224 | 4.34E-04 |
| Gm2035 | 0.001544452 | -8.59E-04 | -0.106668169 | -4.02E-04 | 2.37E-05 |
| Gm2046 | 0.001434175 | -0.001025048 | -0.093631148 | -3.97E-04 | 9.03E-05 |
| Gm2056 | 0.001552841 | -6.49E-04 | -0.11314788 | -4.64E-04 | 1.18E-04 |
| Gm20767 | 5.78E-04 | -3.35E-04 | -0.046638029 | -9.79E-05 | 5.15E-05 |
| Gm21761 | 1.24E-04 | -0.001280855 | -0.324317811 | 6.80E-04 | -8.75E-04 |
| Gm21818 | 0.001029098 | -3.71E-04 | -0.116971594 | -5.00E-04 | 2.37E-04 |
| Gm26710 | -3.97E-06 | -1.37E-04 | -9.08E-04 | 5.89E-04 | -6.69E-05 |
| Gm26737 | 0.003795886 | -0.013865136 | -0.009819154 | -0.005177669 | -0.001232345 |
| Gm26782 | -8.36E-05 | -2.04E-04 | -0.001184615 | 7.85E-04 | -1.07E-04 |
| Gm26870 | -0.048371807 | -0.032894724 | -0.201867714 | -0.067008097 | -0.035084416 |
| Gm26917 | -0.008628551 | -0.035926995 | -0.026976277 | -0.087188205 | -0.026586035 |
| Gm27167 | 3.43E-05 | -1.99E-05 | -0.029962262 | 9.83E-05 | -1.37E-05 |
| Gm29666 | 0.005670184 | -0.018087353 | -0.006227679 | -0.011855667 | -0.003908872 |
| Gm33466 | 6.70E-05 | -9.32E-05 | -0.025738014 | 2.71E-05 | 1.97E-05 |
| Gm36266 | -3.78E-05 | -1.07E-04 | -7.39E-04 | 5.68E-04 | -8.24E-05 |
| Gm37305 | -1.14E-04 | -8.90E-04 | -0.001387111 | 4.54E-04 | -5.34E-04 |
| Gm4027 | 0.001679924 | -0.001192756 | -0.119302968 | -1.24E-04 | -1.09E-04 |
| Gm42418 | 0.006309766 | -0.145318134 | -0.031373339 | -0.042889703 | 0.002727473 |
| Gm42637 | -0.011605206 | 0.007144914 | 0.001258647 | 0.020072541 | -0.006484101 |
| Gm428 | -8.34E-06 | -2.81E-05 | -0.001794996 | 3.40E-04 | -1.27E-05 |
| Gm4340 | 9.41E-04 | -8.59E-04 | -0.066368178 | -2.20E-04 | 3.02E-05 |
| Gm43409 | 0.001892272 | -0.010163325 | -0.003060923 | -0.006964956 | -0.001449 |
| Gm4349 | -7.32E-04 | 4.49E-04 | -0.011768032 | -0.001216676 | -3.70E-04 |
| Gm5039 | 0.001549883 | -9.40E-04 | -0.10736644 | -4.41E-04 | -1.05E-05 |
| Gm5662 | 0.002075387 | 0.005085369 | -0.303546775 | -0.003899683 | 0.00289577 |
| Gm7120 | 0.001516409 | -0.006307339 | -0.011168227 | -0.003324849 | -0.001167514 |
| Gm7325 | -0.033693534 | -0.009986791 | 0.028811975 | -0.129036505 | -0.024066847 |
| Gm8300 | 0.001180977 | 0.003783357 | -0.315622306 | -0.00329574 | 0.001703128 |
| Gm8994 | 0.001559202 | -4.71E-04 | -0.170232309 | -1.22E-04 | 7.12E-04 |
| Gm9112 | -1.13E-04 | 8.87E-05 | -0.001091533 | 5.80E-04 | -4.04E-05 |
| Gmcl1 | 0.003743149 | -0.004968494 | -4.31E-04 | -0.007809663 | 5.43E-04 |
| Gnas | -0.02644024 | 8.61E-05 | 0.00463048 | 0.124679806 | -0.027764272 |
| Golga3 | 0.017609329 | -0.065818501 | -0.038399674 | -0.047434418 | -6.89E-04 |
| Gpd1l | 4.23E-04 | -0.002179557 | 0.00194458 | -0.004139677 | -0.00131444 |

| | | | | | |
|---|---|---|---|---|---|
| Gprc5a | 1.43E-04 | -3.36E-04 | -2.96E-04 | 0.013384502 | -5.87E-04 |
| Gprc5c | -6.79E-05 | -5.53E-05 | -0.001321806 | 0.001577018 | -8.32E-05 |
| Gpx3 | -0.059145077 | 0.037090391 | 0.003871765 | 0.026520426 | -0.048114678 |
| Gpx4 | 0.049080733 | 0.020945682 | -0.060037207 | -0.142074475 | 0.082432351 |
| Grb10 | 0.006268288 | -0.025809465 | 0.005724258 | 0.049618869 | -0.002620839 |
| Grn | -0.043633882 | 0.038688917 | -0.012531061 | -0.002380122 | -0.012617982 |
| Gsn | -0.002754535 | 0.001102072 | 2.55E-04 | 0.028539585 | -0.003046192 |
| Gsta4 | 0.065552187 | 0.100475391 | 0.034164666 | -0.071561843 | 0.114921804 |
| Gstm1 | -0.004104427 | 0.0338219 | 0.008567815 | 0.014464409 | 0.023985567 |
| Gsto1 | -0.020699352 | 0.021062617 | 0.002241935 | 0.060763876 | -0.009394321 |
| Gtsf1l | -0.010177353 | -0.043981798 | 0.037634535 | -0.149164182 | -0.018284859 |
| Guca1a | -0.001317352 | 0.005648774 | -0.047328327 | -0.006610773 | -8.37E-04 |
| H19 | 1.37E-04 | 0.001514035 | -0.001191355 | 0.017758511 | 0.001323573 |
| H1f0 | -0.043444265 | 0.024869088 | -0.030456083 | 0.011997489 | -0.015002011 |
| H2-D1 | -0.001436822 | 0.006240102 | -0.00422713 | 0.021240581 | -0.003529319 |
| H2-T23 | 0.00165844 | -0.003714318 | -0.003415006 | -0.002326803 | -7.70E-04 |
| H2afx | -0.019625377 | -0.008498125 | -0.046000404 | -0.14422322 | 0.005903131 |
| H2afy | -0.034327621 | 0.033212627 | 0.010702678 | 0.010072446 | 0.004995494 |
| H3f3b | -0.008073847 | -0.013378068 | -0.025110712 | 0.056244883 | -0.019505153 |
| Hand1 | -1.05E-04 | 6.74E-05 | -8.33E-04 | 0.001585217 | -1.29E-04 |
| Hat1 | -0.004469954 | 0.117902815 | 0.03440926 | -0.034486201 | 0.021256657 |
| Hbegf | -0.003598156 | -0.008759868 | 0.005555942 | 0.028965254 | -0.013958616 |
| Hck | 0.019281404 | -7.77E-04 | 0.002449918 | -0.027746485 | 0.020792285 |
| Hdc | -8.79E-05 | 7.32E-05 | -0.001518002 | 5.91E-04 | -6.93E-05 |
| Herpud1 | 0.002894335 | -0.001382962 | -0.005617303 | 0.012313442 | 0.009740679 |
| Hes1 | -0.009284146 | 0.003444164 | 0.003157892 | 0.00645871 | -0.010558028 |
| Hes6 | -0.011916534 | 0.002647025 | 0.006956105 | 0.045780951 | -0.015236981 |
| Hexa | -0.028245511 | 0.020392295 | 0.001251065 | 0.017808653 | -0.007198755 |
| Hexb | -0.014351274 | 0.027310912 | -0.009594757 | -0.018444238 | 0.005198014 |
| Hist1h1a | 0.015481442 | -0.056252233 | -0.005608745 | -0.016923621 | -0.006775839 |
| Hist1h1b | 0.064062664 | -0.185105661 | 0.038754887 | -0.033415342 | -0.016826982 |
| Hist1h1e | 0.00485218 | -0.126583793 | 0.018461544 | 0.041800408 | -0.023596863 |
| Hist1h2aa | -1.10E-05 | -1.58E-04 | -0.001068486 | 9.29E-04 | -6.80E-05 |
| Hist1h2ae | 0.027931222 | -0.054981488 | 0.001300678 | -0.008187614 | -4.64E-04 |
| Hist1h2ap | 0.032154737 | -0.02691278 | 0.00444087 | -0.010091435 | -0.009922881 |
| Hist1h2bc | -0.009476675 | 0.015737616 | -0.040654624 | -0.01702401 | -0.024702514 |
| Hist1h3c | -1.68E-04 | -0.001953258 | -0.022463785 | 2.80E-04 | 3.55E-04 |
| Hist1h3d | -1.98E-05 | -1.37E-04 | -0.010259774 | 6.34E-04 | -1.61E-04 |

| | | | | | |
|---|---|---|---|---|---|
| Hist1h4i | -0.001313371 | -0.00316427 | -0.082329336 | -0.004846332 | -0.007704536 |
| Hist1h4j | -7.64E-05 | -3.07E-05 | -0.002316947 | 7.58E-04 | -1.22E-04 |
| Hist2h3c1 | -3.17E-05 | 3.55E-05 | -0.001744998 | 4.00E-04 | -7.42E-05 |
| Hist3h2a | -0.004703836 | 0.003463334 | -0.001004246 | -0.001278379 | -0.004475031 |
| Hist3h2ba | -0.002059292 | 2.27E-04 | -0.01983069 | 0.005520341 | -0.002532476 |
| Hmces | 0.063553348 | -0.007531235 | 0.066383834 | -0.200492651 | 0.034617147 |
| Hmga2 | -0.002339191 | -0.007566669 | 4.97E-04 | 0.042552794 | -0.005257593 |
| Hmgb3 | 0.014942031 | -0.004215895 | 0.002666418 | 0.054640007 | 0.011420899 |
| Hmgcr | 0.024477735 | -0.051996331 | 0.00558123 | 0.03606006 | -0.042784383 |
| Hmgn3 | -0.003110238 | -3.35E-04 | -0.055869518 | 0.044010291 | -0.003875943 |
| Hmgn5 | -0.009835818 | 0.007378859 | -0.02464618 | 0.040849864 | 0.002985787 |
| Hmmr | -0.046006621 | 0.004140579 | -0.044823236 | -0.046304198 | -0.012948049 |
| Hmox1 | 0.061353944 | -0.001346621 | 0.023404713 | 0.083690324 | 0.020044198 |
| Hnrnpa1 | 0.061170224 | -0.033787405 | 0.01081182 | 0.094396448 | 0.011273987 |
| Hormad1 | -0.009264584 | 0.008450698 | -0.001152246 | -0.010439933 | -0.00440209 |
| Hoxa7 | -2.97E-04 | 1.26E-04 | -0.001399197 | 0.001289343 | -3.27E-04 |
| Hpcal1 | -0.0339477 | 0.026875318 | 0.005832616 | 0.019091925 | -0.029492456 |
| Hsd17b14 | 0.011089337 | -0.044735092 | 0.106051413 | -0.346472531 | -0.044917574 |
| Hsf2 | 0.02605336 | -0.053801228 | -3.12E-06 | 0.01317644 | -0.055499634 |
| Hsp90aa1 | 0.040033775 | 0.103634877 | -0.004815912 | 0.024185591 | 0.123992784 |
| Hsp90ab1 | -0.062028004 | 0.115855422 | -0.024276304 | -0.034291196 | 0.012435059 |
| Hsp90b1 | 0.057612128 | -0.020935227 | 0.011896731 | 0.04975851 | 0.069348113 |
| Hspa1a | -1.75E-04 | 1.50E-04 | -0.002073984 | 0.001774532 | -2.43E-04 |
| Hspa5 | 0.076528532 | 0.003503769 | 0.019792154 | 0.070144814 | 0.091377908 |
| Hspa8 | 0.012739096 | 0.189514438 | -0.045972985 | 0.021063916 | 0.075811342 |
| Hspb1 | -0.016773864 | 0.100173392 | -0.007846261 | -0.169174187 | -3.21E-04 |
| Hspb2 | -0.003230929 | 0.002432434 | 2.70E-05 | 0.013752853 | -0.002830931 |
| Htra1 | 0.001318056 | 0.004833675 | 0.00481156 | 0.008540809 | 0.006121752 |
| Icam1 | 0.00630797 | 0.00703989 | 0.008989258 | -0.013981006 | 0.003087132 |
| Id1 | 0.040065972 | -0.04919378 | -0.033985566 | 0.111454095 | -0.018227382 |
| Id2 | 0.022338475 | -0.020096399 | -0.029054825 | 0.148246917 | 0.006687978 |
| Id3 | 0.096115824 | -0.026196369 | -0.078360946 | 0.050657793 | 0.047180667 |
| Ier3 | 0.001120546 | 0.001575715 | -3.56E-04 | 0.020101386 | 0.00213782 |
| Ier5 | -7.88E-04 | 1.99E-04 | -0.001384854 | 0.013608566 | -0.001971712 |
| Ifi30 | -0.006581221 | 0.014093334 | 0.004032763 | 0.015225148 | -6.38E-04 |
| Ifitm1 | 0.031893569 | 0.046594319 | 0.048821462 | -0.066936477 | 0.158903671 |
| Ifitm2 | 0.009444954 | 0.069888251 | 0.062751191 | -0.215822579 | 0.068095611 |
| Ifitm3 | -0.02227293 | 0.021247014 | 0.032649632 | 0.094260389 | 0.034847059 |

| | | | | | |
|---|---|---|---|---|---|
| Ifrd1 | 0.042927653 | -0.060322708 | 0.001797151 | 0.116655365 | 0.025717319 |
| Igf2 | -0.001106734 | -0.001035201 | -1.01E-04 | 0.03730004 | -0.001021383 |
| Igfbp2 | 0.048492835 | 0.053714694 | -0.070151524 | 0.01119684 | 0.089111915 |
| Igfbp3 | 0.010198626 | -0.007622744 | -0.051742968 | 0.059461642 | 0.002757366 |
| Igfbp4 | 2.02E-04 | -0.001168631 | -1.74E-04 | 0.007241916 | -9.25E-04 |
| Igsf23 | -0.004495851 | 8.05E-04 | 0.005000927 | -0.025721906 | -0.002831072 |
| Il23a | -1.23E-04 | -1.06E-04 | -0.001188404 | 0.001151063 | -7.41E-05 |
| Impact | 0.001442159 | -0.061812266 | 0.015261519 | 0.104523722 | -0.063156827 |
| Ina | 0.066211805 | -0.052238787 | 0.033192738 | 0.008516785 | 0.034721849 |
| Insig1 | 0.013572253 | -0.025435021 | -0.003064399 | 0.010946805 | -0.027900189 |
| Irf1 | 0.001811446 | -0.0034823 | 0.001531166 | 0.024622773 | 1.26E-04 |
| Irgm1 | -4.88E-05 | -7.36E-04 | -0.001366376 | 0.003105408 | -3.79E-04 |
| Itga3 | 0.006718506 | -0.023619893 | 0.004051268 | 0.061864914 | -0.004849552 |
| Itgb1 | 0.020029311 | -0.061455753 | 0.012513786 | 0.10528726 | -1.54E-04 |
| Itm2a | -0.017393764 | 0.01706275 | -0.007909486 | 0.012558346 | -0.008229278 |
| Jam2 | -0.035622967 | 0.033153645 | -0.039081728 | -0.208039136 | 0.008395053 |
| Jarid2 | -0.00578902 | -0.178335454 | 0.032915926 | -0.10858667 | -0.018078995 |
| Jun | -0.009782796 | 0.001643707 | -0.017005339 | 0.03960419 | -0.008617845 |
| Junb | -0.007204448 | -0.012485814 | 0.022476019 | 0.056498236 | -0.016513919 |
| Jund | -0.042146884 | -0.011434556 | 0.024610359 | 0.105489615 | -0.043178208 |
| Kcnk1 | 0.012389903 | -0.00948012 | 0.003868449 | 0.062472283 | 0.007928033 |
| Kctd12 | -1.49E-04 | 2.68E-05 | -0.001440394 | 9.90E-04 | -1.31E-04 |
| Kdelr3 | -0.011338847 | 0.006649334 | 0.005593889 | -0.016479096 | -0.014439081 |
| Kdm5b | -0.047258927 | -0.024140177 | -0.017706916 | -0.158060359 | 0.00834516 |
| Kdm6b | 0.009760064 | -0.032905533 | 0.005703568 | 0.002663588 | -0.048290782 |
| Khdc3 | 9.67E-04 | 0.028619104 | 0.01104553 | -0.08141985 | 0.016683645 |
| Kif1a | -0.002951886 | -0.002984681 | 6.90E-04 | 0.025369345 | -0.002773074 |
| Kif21a | 9.94E-04 | -0.005826411 | 0.001364508 | 0.020831772 | -0.002059122 |
| Klf2 | 0.008122799 | 0.035622749 | 0.015861479 | -0.307922095 | 0.125566655 |
| Klf4 | -0.057337324 | 0.02829506 | 0.028441218 | -0.18353764 | -0.023984693 |
| Klf5 | -0.012202198 | -0.010984705 | 0.010210949 | -0.022813991 | -9.19E-04 |
| Klf6 | -0.021867614 | -0.011608955 | 0.002042234 | 0.114326704 | -0.00210744 |
| Klf9 | -0.064598374 | -7.57E-04 | -0.013922344 | -0.063991781 | -0.01069032 |
| Klhl13 | 0.046445627 | 0.045251506 | -0.032789733 | 0.076959356 | 0.110808284 |
| Kpna7 | -7.43E-06 | -3.63E-06 | -0.00101311 | 6.53E-04 | 2.81E-05 |
| Kras | 0.011196368 | -0.073045032 | 0.028296469 | 0.145576114 | -0.016694108 |
| Krt17 | 0.00268999 | 0.001718747 | 0.003474725 | 0.001693131 | 0.003114328 |
| Krt18 | -0.002980509 | 0.004904913 | -0.002480214 | 0.171910917 | -0.011846927 |

| | | | | | |
|---|---|---|---|---|---|
| Krt19 | -0.004916974 | 0.001510955 | 0.001600314 | 0.057531649 | -0.006035514 |
| Krt42 | -0.001090608 | 0.014615946 | 0.015313514 | -0.055519108 | 0.008684051 |
| Krt7 | -0.002919847 | 0.00157334 | 6.00E-04 | 0.024293101 | -0.003001104 |
| Krt8 | -0.007882698 | 0.003019604 | -0.003074924 | 0.155167219 | -0.014985815 |
| L1td1 | 0.0693294 | -0.160030121 | 0.073180725 | -0.043800132 | 0.034376989 |
| Lama1 | -0.002816016 | -0.00526691 | 0.002512223 | -0.009222626 | -0.004344234 |
| Lamb1 | 0.004354365 | -0.013980449 | -0.003288048 | 0.028755435 | 0.003826097 |
| Ldha | 0.105078206 | 0.066167015 | 0.043405192 | 0.048218227 | 0.088851786 |
| Ldhb | -0.016269053 | 0.120017739 | 0.009740181 | -0.067205694 | 0.072225293 |
| Ldoc1 | 6.03E-04 | -4.37E-04 | -0.002050912 | 0.004655214 | 3.50E-04 |
| Lef1 | 0.015216636 | -0.046828357 | 0.014431509 | 0.017416098 | -0.004533445 |
| Lefty1 | -0.02708893 | 0.01521308 | 0.017654762 | 0.184881342 | -0.006265126 |
| Lefty2 | -0.008429617 | -1.77E-04 | 0.009629743 | 0.075761382 | -0.004611015 |
| Lepr | 0.012620998 | -0.03822612 | -0.027655292 | -0.028369726 | -0.006545408 |
| Lgals1 | -0.002913792 | -0.002916448 | 0.02394044 | 0.239228925 | -0.010333188 |
| Lgals3 | -0.031610612 | 0.018309868 | 3.56E-04 | -0.080949001 | -0.027170765 |
| Lgals4 | -1.78E-04 | 0.001517704 | -0.079685941 | 0.001179022 | -4.58E-04 |
| Lgmn | 1.75E-04 | 0.018291372 | 0.002196038 | 0.012280877 | 0.01132257 |
| Lgr4 | 0.019292961 | -0.056497055 | 0.005496263 | 0.014404797 | -0.052293247 |
| Limch1 | -6.69E-04 | 7.82E-04 | -0.053491807 | 9.93E-04 | -1.41E-04 |
| Limk2 | 0.019883691 | -0.06279314 | -0.018201957 | -0.015907789 | -0.009326561 |
| Lmo4 | 0.010125191 | -0.024084184 | -0.015159741 | 0.054703877 | -0.005344354 |
| Lmo7 | 0.002039972 | 5.64E-04 | 9.58E-04 | -0.002562058 | 0.005918777 |
| Lmx1a | -3.18E-05 | -7.87E-06 | -0.002703354 | 4.53E-04 | -2.16E-05 |
| Lncenc1 | 0.006936803 | -0.039712734 | -0.002566082 | -0.127924613 | 0.010871613 |
| Lonp2 | -0.00442493 | 0.008353421 | -0.106099391 | 0.006641104 | 0.001233349 |
| Lor | 6.67E-04 | -0.001097469 | -0.00319076 | 0.008326018 | -3.27E-04 |
| Loxl2 | 0.024081635 | -0.064187591 | -0.018651662 | -0.024101438 | -0.006659373 |
| Lpar1 | 0.004694242 | 4.36E-04 | -0.006247896 | -0.016719301 | 0.006856623 |
| Lpar4 | -0.003399593 | -0.006495016 | -4.62E-04 | 0.026953385 | -0.006142586 |
| Lpar6 | -0.008843043 | 0.001143145 | -0.003347976 | 0.033403875 | -0.009517438 |
| Lpl | 9.31E-04 | -6.64E-05 | -0.003994969 | 0.00419798 | 0.001162876 |
| Lpp | -0.003183668 | -0.013949157 | 0.003166762 | 0.041251556 | -0.005208907 |
| Lrp2 | 0.008285361 | -0.023846038 | 0.00170884 | 0.030929483 | 0.0017057 |
| Lrpap1 | 0.030010745 | 0.049135124 | -0.018775201 | 0.038380018 | 0.079965519 |
| Lrrc8a | 0.011651537 | -0.034420484 | -0.002320984 | 0.032682445 | -0.041489026 |
| Lrrn4 | 6.97E-04 | -9.86E-04 | -0.00884764 | 0.007466712 | 1.38E-04 |
| Ltb | 0.002955557 | -0.007726991 | 9.17E-04 | 0.039099967 | -0.002019016 |

| | | | | | |
|---|---|---|---|---|---|
| Ly6a | -0.001017355 | 0.013027742 | -0.002060189 | -0.005084721 | 0.016410421 |
| Ly6g6e | -0.017432326 | 0.018658609 | 0.002819784 | -0.018690375 | -5.23E-04 |
| Ly6k | -4.43E-04 | 5.67E-04 | -0.004510612 | -2.97E-04 | -3.54E-04 |
| Malat1 | -0.279448406 | -0.147444223 | 0.093435204 | 0.173827295 | -0.021093659 |
| Manba | -0.013986797 | -0.008680398 | 0.010080508 | -0.107235958 | -0.029387813 |
| Manf | 0.071189473 | 0.014602145 | 0.00697793 | 0.037214226 | 0.058005961 |
| Map1b | 0.007198015 | -0.027305661 | 0.001214677 | 0.033425405 | 0.008240411 |
| Mapt | -0.009655641 | 3.07E-04 | 0.00971489 | -0.04595987 | -0.00856217 |
| Marcks | -0.027597535 | -0.009584052 | 0.025857711 | 0.074802113 | -0.027544963 |
| Marcksl1 | 0.010602222 | -0.014656572 | 0.047711419 | 0.069583481 | -0.011188035 |
| Marfl | 0.012099078 | -0.046681972 | 0.003119679 | 0.003475733 | -0.063438731 |
| Marveld1 | -0.002902218 | 0.001134494 | -0.026238245 | 8.53E-04 | -0.001710972 |
| Mbd5 | -3.15E-04 | -6.94E-04 | -0.014495716 | 0.001331401 | -4.26E-04 |
| Mbnl3 | 7.02E-04 | -9.01E-04 | -6.54E-04 | 0.01163144 | -5.89E-04 |
| Mdk | -0.017799147 | -0.044851346 | 0.010455647 | 0.094610832 | -0.048776816 |
| Mdm2 | -0.13549539 | 0.098486435 | -0.007729994 | 0.050519678 | -0.119054953 |
| Mecom | 5.42E-05 | 9.47E-05 | -0.036160386 | 9.44E-04 | 1.63E-04 |
| Meg3 | -0.034614851 | -0.18395029 | 0.075182549 | 0.117896147 | -0.058680528 |
| Mest | 0.055823298 | -0.066856316 | 0.002526378 | 0.144922384 | 0.034313628 |
| Mfge8 | -0.029252526 | 0.049960181 | -6.86E-04 | -0.002929866 | 0.006035723 |
| Mfsd1 | -0.011243933 | 0.006377574 | -0.009365582 | 0.014300776 | -0.001905504 |
| Mgll | -1.04E-04 | 1.21E-04 | -0.036858247 | 4.92E-04 | 6.77E-06 |
| Micu1 | -0.003179157 | 0.00207785 | -0.024829364 | 0.001813925 | -0.00212678 |
| Mif | 0.053609472 | -0.003616128 | 0.062402516 | 0.054832779 | 0.052932426 |
| Mis18bp1 | -0.03140488 | -0.037529006 | -0.072244253 | -0.054439107 | 0.004125188 |
| Mkrn1 | 0.01194717 | -0.064751835 | -0.012517709 | -0.165092594 | -0.048841611 |
| Mme | 0.007481995 | -0.008561705 | -0.00682748 | 0.011310794 | 0.013268678 |
| Mobp | -0.005104031 | 0.007457638 | -7.49E-04 | -0.009792575 | 0.001949643 |
| Morc1 | 0.002016087 | -0.001477515 | 0.031910235 | -0.170226748 | 0.008687144 |
| Morc4 | 2.69E-04 | -0.002536429 | -0.003384715 | 0.016316462 | -6.60E-04 |
| Mpzl1 | 0.002513566 | 0.013817346 | -0.046569409 | 0.031635926 | 0.005665422 |
| Mras | 0.002265169 | -0.028642506 | 0.022862827 | -0.095544693 | -0.012597354 |
| Mreg | -5.81E-04 | 0.009955658 | -0.072187041 | -0.050911294 | -0.012870218 |
| Mrpl9 | 0.015929406 | -0.021631762 | -0.001427986 | 0.009960488 | 0.005733748 |
| Ms4a10 | -0.019586487 | 0.014178559 | 0.001086156 | -0.003023907 | -0.01130517 |
| Msc | -0.003622729 | -0.010812182 | 0.01428403 | -0.054901403 | 0.001893051 |
| Msmo1 | 0.010892868 | 0.008194122 | 0.010504252 | 0.018413688 | 0.01625965 |
| Msx1 | 1.48E-04 | -1.90E-04 | -0.002162973 | 0.003483059 | -2.90E-04 |

| | | | | | |
|---|---|---|---|---|---|
| Msx2 | -0.00248295 | 0.005301922 | -0.011661012 | 0.007866211 | 0.001500526 |
| mt-Atp6 | 0.003177401 | -0.131681907 | 0.005973945 | -0.185512095 | 3.31E-04 |
| mt-Co1 | -0.031802819 | -0.140797854 | -0.023787161 | -0.141162489 | -0.0429893 |
| mt-Co2 | -0.046539888 | -0.15960579 | -0.007447199 | -0.187360931 | -0.028969444 |
| mt-Co3 | -0.027583513 | -0.082045197 | -0.003349011 | -0.180695569 | -0.011024762 |
| mt-Cytb | 0.001786661 | -0.089233464 | 0.008360164 | -0.174961394 | 0.016060353 |
| mt-Nd1 | 0.039169866 | -0.114429133 | 0.021088817 | -0.183941797 | 0.037854954 |
| mt-Nd2 | -0.023355082 | -0.133786102 | 0.010605147 | -0.141818188 | -0.029212122 |
| mt-Nd4 | -0.03969539 | -0.111213792 | -0.006141267 | -0.179387729 | -0.027555985 |
| Mt1 | 0.033603601 | -0.057090086 | 0.048552914 | 0.025042508 | 0.005734125 |
| Mt2 | 0.039408668 | -0.041825396 | 0.036118244 | -0.006745423 | 0.02027861 |
| Mtf2 | 0.006378273 | -0.019924715 | -0.06491704 | -0.173754388 | 0.024113626 |
| Mthfd2 | 0.087404661 | -0.046644878 | 0.012088697 | 0.037497392 | 0.054573534 |
| Muc3 | -9.12E-05 | -2.49E-05 | -9.95E-04 | 5.59E-04 | -8.02E-05 |
| Myc | 0.023370936 | -0.020629784 | 0.010825168 | 0.017769745 | 6.53E-04 |
| Mycn | 0.067411073 | -0.044042419 | 0.044545734 | 0.087931627 | 0.075407746 |
| Myef2 | 0.076265835 | -0.168188017 | 0.017395789 | 0.053682202 | -0.162611084 |
| Myh13 | -0.001528134 | 0.002047328 | -0.004738197 | -0.00456261 | 4.66E-04 |
| Myl4 | -0.003342614 | 1.40E-04 | -0.001455449 | -0.008208131 | -0.004188803 |
| Myl6 | 0.024229041 | 0.041755942 | -0.005965539 | 0.086618219 | 0.049565419 |
| Myl9 | 0.038173143 | -0.015026056 | 0.00655825 | 0.082022641 | 0.033220232 |
| Mylpf | 0.00762264 | 0.016582394 | -0.005905894 | -0.330180946 | 0.017515993 |
| Myo1f | -0.011030875 | -0.001111674 | 0.008116494 | -0.049552705 | -0.014188287 |
| Nab2 | -9.02E-04 | -0.003292973 | 0.003788052 | 2.18E-04 | -0.001946544 |
| Nanog | 0.018688981 | 0.056836233 | -0.008471113 | -0.120264021 | 0.079350589 |
| Napsa | 0.002106286 | 0.002385003 | 0.003459039 | -0.010069528 | 0.002921363 |
| Ncoa3 | 0.054139592 | -0.145154274 | 0.016198142 | -0.053348413 | -0.154528453 |
| Ndrg1 | -0.002589028 | 3.61E-04 | -0.001381648 | 0.005073681 | -0.001648713 |
| Ndufa4l2 | -0.009347867 | 0.019387993 | -0.004908015 | -0.008423932 | -0.0037948 |
| Neat1 | -0.033189752 | -0.025853539 | 0.00393266 | 0.023543036 | -0.015895944 |
| Nefh | 0.025385395 | -0.026736913 | -0.00378565 | -0.027549746 | 0.007290907 |
| Nefl | 0.012092518 | -0.008749394 | 0.003422366 | 0.019686029 | 0.005906706 |
| Nefm | 2.39E-05 | -3.28E-04 | -9.46E-04 | 0.002264941 | -9.34E-05 |
| Nelfa | 0.011562373 | -0.006992268 | -0.182897903 | -0.005356775 | 1.74E-04 |
| Nes | 0.003465965 | -0.013198811 | 1.40E-05 | 0.058100284 | -0.003279728 |
| Neurod1 | -0.026571582 | 0.015595761 | 0.002418989 | 0.002174998 | -0.017704936 |
| Neurog3 | 6.47E-05 | -5.76E-05 | -0.001051491 | 6.86E-04 | 2.56E-05 |
| Nexn | 0.006854108 | -0.005338865 | 0.001274971 | 0.008301595 | 0.006757808 |

| | | | | | |
|---|---|---|---|---|---|
| Nfatc2ip | 0.008654283 | -0.004753409 | 0.005003053 | -0.151341233 | 0.01532826 |
| Nfkbia | -0.028951272 | 0.034448971 | -0.029404953 | 0.005198754 | -0.013678806 |
| Ngb | -3.98E-05 | -8.38E-05 | -0.001043445 | 8.20E-04 | -8.51E-05 |
| Ngfrap1 | 0.048187073 | 0.00631662 | 0.027065202 | 0.052919149 | 0.055472698 |
| Nip7 | 0.11387971 | -0.116074804 | 0.019283563 | -0.123949609 | 0.02526126 |
| Nkx2-9 | -0.030433385 | 0.019275775 | 0.001609255 | -0.004328027 | -0.025809169 |
| Nkx6-2 | -0.016433635 | 8.91E-04 | 0.003086592 | -0.00703547 | -0.025059342 |
| Nmrk1 | 8.94E-04 | -0.009448927 | -0.005038132 | -0.002738313 | -0.001998297 |
| Nnat | 3.77E-04 | -0.013131531 | 0.004217984 | 0.058933167 | -0.008536158 |
| Noct | 0.04423333 | -0.033478622 | 0.005857229 | 0.004174658 | 0.007174826 |
| Nodal | -0.03129867 | 0.023674086 | 0.001045456 | -0.011709675 | -0.016905471 |
| Nop56 | 0.058848103 | 0.018406202 | -0.015415849 | 0.019113126 | 0.104539997 |
| Notch2 | 0.016302808 | -0.032451641 | 0.001413942 | 0.025744394 | -0.034391706 |
| Nptxr | -1.54E-04 | -8.08E-05 | -0.001092939 | 0.001536581 | -3.12E-04 |
| Npy | -1.44E-04 | 1.18E-04 | -0.001364755 | 6.60E-04 | -1.42E-06 |
| Nr0b1 | 0.037417734 | 0.0233345 | 0.028155083 | -0.150992843 | 0.030194427 |
| Nrk | -3.09E-04 | 1.78E-04 | -9.19E-04 | 0.00226351 | -2.22E-04 |
| Nrp2 | -0.001011048 | -0.003917209 | 0.003096471 | 0.012987042 | 0.002853602 |
| Nudt4 | 0.065859193 | -0.065587955 | -0.044712344 | 0.007821078 | 0.023398883 |
| Nup62cl | 0.01261397 | 0.022636276 | -0.003475617 | 0.022333272 | 0.023147364 |
| Nupr1 | 0.009942 | 0.01078577 | 0.008430029 | -0.003685962 | 0.023428487 |
| Ooep | -0.007366435 | 0.066640139 | 0.018363075 | -0.125079248 | 0.035893591 |
| Os9 | -0.00198138 | 0.001010564 | -0.035754337 | 0.001886776 | -7.37E-04 |
| Otx2 | 0.001965149 | -0.001240393 | 0.007625131 | -0.002152088 | 0.004949408 |
| P2rx4 | -0.002693713 | -0.002268247 | -1.24E-04 | 0.004552307 | -0.003283805 |
| P2rx7 | -0.001405822 | 7.47E-04 | -0.001105004 | 0.001051643 | -0.0010008 |
| P3h3 | -0.001271302 | -0.007809642 | -0.003065571 | -0.008690969 | -0.004587411 |
| P3h4 | 0.024921767 | -0.025505157 | 0.0132048 | 0.008935724 | 0.004087212 |
| P4ha2 | -0.003196753 | 0.004073027 | -0.030754916 | 0.010441767 | -0.002352356 |
| Parm1 | -0.001419063 | 0.001259753 | -0.001198214 | 0.001442999 | 1.89E-04 |
| Parp1 | 0.046715437 | -0.16797461 | -0.020898591 | -0.163572803 | -0.007998232 |
| Pawr | 0.008130684 | -0.01240657 | 0.007230153 | 0.049025798 | 0.010729179 |
| Pax6 | 0.011980882 | -0.011328142 | 0.005606282 | -0.011516659 | 0.007872575 |
| Pcgf5 | 0.003634731 | -4.36E-04 | -0.012572193 | 0.047488577 | 0.004315397 |
| Pdcd4 | -0.00789168 | -0.025852432 | -0.007503915 | -0.004001627 | -0.072177537 |
| Pdgfa | -0.012398769 | -0.027079324 | 0.021827027 | -0.063898847 | -0.019116512 |
| Pdgfrl | -8.26E-04 | 4.69E-04 | -0.053583074 | 0.003126072 | -4.97E-04 |
| Pdia3 | 0.005369373 | 0.017888703 | -0.004200175 | 0.051094224 | 0.024729767 |

| | | | | | |
|---|---|---|---|---|---|
| Pdia6 | 0.050939568 | 0.067718363 | 0.018118427 | 0.052659041 | 0.083918056 |
| Pdlim3 | -1.76E-04 | 2.04E-04 | -0.003512829 | 0.001354037 | -5.54E-05 |
| Pdlim4 | -6.78E-04 | 0.002376229 | -0.001473164 | 0.003117302 | 8.70E-04 |
| Pdlim7 | 0.020585711 | -0.02158738 | -0.005269746 | 0.15964993 | 0.004435017 |
| Pdpn | 0.014262618 | -0.008320707 | -0.004190804 | 0.029969307 | 0.006418388 |
| Peg10 | 0.014211969 | -0.064341817 | 0.005822721 | 0.10274513 | -0.014651889 |
| Peg3 | -0.004358451 | -0.018175785 | -1.67E-04 | 0.034414603 | -0.001898183 |
| Perp | -0.0932979 | 0.03752984 | 0.011761335 | 0.144283385 | -0.084453512 |
| Pfas | 0.06476031 | -0.154837931 | 0.016723907 | 0.004444182 | -0.143923056 |
| Pfkp | 0.018230176 | 0.033052037 | -0.002268841 | -0.07208427 | 0.03837191 |
| Pfn1 | 0.085239224 | -0.065621214 | 0.072980703 | 0.105144885 | 0.042757368 |
| Pga5 | -0.003691378 | 0.007133614 | 9.78E-04 | -0.007490221 | 0.002818068 |
| Phgdh | 0.083580129 | 0.033859879 | -0.007666345 | 0.118330375 | 0.091499574 |
| Phkg1 | -7.65E-05 | -0.004851624 | -5.51E-04 | -0.007518374 | -0.002191424 |
| Phlda1 | -0.017716848 | -0.003066676 | 3.28E-04 | 0.067060241 | -0.028876969 |
| Phlda2 | 0.059807473 | -0.065739438 | 0.015924514 | 0.091412167 | 0.027658565 |
| Phlda3 | -0.156351078 | 0.112591958 | 0.014032296 | 0.083272528 | -0.150804628 |
| Phldb2 | -3.74E-04 | -0.001557033 | -0.007006146 | 0.017095287 | -0.001861342 |
| Pim1 | 0.042080292 | -0.017677341 | -0.014440433 | 0.135370357 | 0.029781873 |
| Pim2 | 0.043183702 | -0.061993812 | 0.008671562 | 0.154306176 | 0.011228147 |
| Pitx2 | 2.72E-04 | -0.003519821 | 3.05E-04 | 0.026479618 | -0.002427891 |
| Pkp2 | -0.008570319 | -6.15E-05 | 0.009903538 | 0.004402135 | -0.009734447 |
| Plac8 | 0.021314505 | -0.00489952 | 0.002697094 | 0.026092838 | 0.014144945 |
| Platr3 | 0.008229817 | -0.034911261 | -0.01139194 | -0.123509221 | -0.009300424 |
| Platr31 | -8.92E-05 | 1.04E-04 | -0.097282535 | -3.08E-05 | -1.05E-04 |
| Plaur | 0.001702406 | -0.014375227 | 0.005672787 | 0.076877213 | -0.009382456 |
| Plet1 | -7.47E-04 | 5.76E-04 | -8.16E-04 | 0.00441249 | -5.32E-04 |
| Plin2 | -0.010407313 | 0.013807161 | -0.011091295 | 0.068974875 | -0.003919518 |
| Plk2 | -0.094190637 | 0.069194166 | -0.01911805 | 0.046742238 | -0.078236084 |
| Plk3 | -0.001409872 | 0.002207474 | -0.006693098 | 0.018961133 | -0.003843036 |
| Plk4 | 0.068106068 | -0.204552106 | -0.036934095 | -0.124543503 | -0.013710003 |
| Plod2 | -0.028104822 | 0.009768466 | -0.002757666 | 0.021587922 | -0.02447292 |
| Pmaip1 | -0.021713652 | 0.010117411 | -0.037237525 | 0.023244276 | -0.023055701 |
| Pmm1 | -0.118941212 | 0.068876397 | 0.056549966 | -0.016122822 | -0.10291986 |
| Pmp22 | 3.69E-05 | -5.77E-04 | -6.12E-04 | 0.010240177 | -7.75E-04 |
| Pou3f1 | 0.024924201 | -0.063773519 | 0.003318933 | 0.212071578 | -0.010763637 |
| Ppdpf | 0.019645287 | -0.045979657 | 0.021001809 | -0.125119235 | 5.79E-04 |
| Ppig | 0.018267494 | -0.04696874 | -0.084325227 | -0.005687535 | 0.021555108 |

| | | | | | |
|---|---|---|---|---|---|
| Pramel7 | -3.63E-05 | 3.18E-05 | -0.001328859 | 4.64E-04 | -1.72E-05 |
| Prdx1 | 0.090999259 | 0.102578976 | 0.009625363 | 0.034291555 | 0.11891914 |
| Prdx6 | 0.151273324 | 0.022948875 | -0.02512435 | 0.154307814 | 0.162160303 |
| Prex2 | -0.003842595 | -0.007761019 | -0.090860088 | 0.004845453 | 2.42E-04 |
| Prph | -0.004727077 | 0.003031429 | 0.001361769 | 0.016044812 | -0.004206345 |
| Prps1 | 0.067951947 | 0.062895702 | -0.099460969 | 0.007374221 | 0.092581007 |
| Prr13 | -0.013072043 | -0.008890539 | 0.019357494 | -0.016001119 | -0.020239692 |
| Prrc1 | -0.001712161 | 0.011354521 | -0.020761412 | -0.073399484 | 0.00241022 |
| Prss8 | -1.99E-05 | -0.001111835 | 9.79E-04 | 0.02171137 | -0.001293574 |
| Prtg | 3.03E-04 | -0.013661149 | 0.003712767 | 0.055582774 | -0.005875246 |
| Psap | -0.019291258 | 0.03668881 | -0.020065612 | 0.021865415 | 0.011627691 |
| Psmb8 | 1.25E-04 | -3.35E-04 | -0.001140362 | 0.003950877 | -1.50E-04 |
| Psors1c2 | 0.018406771 | -0.010691131 | 0.005942412 | 0.022180466 | 0.012399972 |
| Psrc1 | -0.143519509 | 0.047059449 | 0.025567188 | -0.012807948 | -0.117812792 |
| Ptch1 | -0.002802798 | -0.010159911 | 0.007861222 | -5.78E-04 | 0.00451781 |
| Ptcra | 0.00944601 | -0.031962909 | -0.004120559 | -0.024674661 | -0.010375746 |
| Ptges | 9.38E-04 | 0.001267701 | 0.001026326 | 0.003495016 | -1.30E-04 |
| Ptgr1 | -0.008275283 | 0.021970981 | -0.011398685 | 0.01542607 | 0.006400384 |
| Pth1r | -1.01E-04 | 6.14E-05 | -0.001259953 | 0.002013368 | -2.07E-04 |
| Ptp4a2 | 0.031366001 | -0.117954183 | 0.018700084 | 0.061608144 | -0.125905783 |
| Ptp4a3 | -0.134357685 | 0.067374592 | 0.01689763 | -0.012826827 | -0.109204362 |
| Pttg1 | -0.015572055 | 0.058570836 | -0.056964406 | 0.075870962 | 0.01250702 |
| Pxdc1 | -2.37E-04 | -4.17E-04 | -0.001090518 | 0.00696591 | -6.86E-04 |
| Pycr2 | 0.02720122 | 0.001512636 | 0.034263544 | 0.155656412 | 0.02227133 |
| Rab3il1 | -0.001761406 | 0.00345759 | 1.40E-04 | 0.00342679 | 2.55E-04 |
| Rap2b | -0.040493995 | -0.01720948 | 0.004639087 | 0.053792529 | -0.079144424 |
| Rasgrp2 | 0.142654975 | -0.084145431 | 0.067187957 | -0.098523243 | 0.099713945 |
| Rasl11a | -0.007115677 | 0.011252226 | -0.003726456 | 0.002491248 | -0.012029338 |
| Rassf1 | 0.001880055 | 1.61E-04 | 0.00412675 | 0.040478173 | -0.007863496 |
| Rbm15 | 0.047115154 | -0.10250379 | -0.006951642 | 0.015734234 | -0.112679503 |
| Rbm47 | 0.0039574 | -2.63E-04 | 0.00364418 | -0.002009678 | 0.004685175 |
| Rbms1 | -0.00479094 | -0.024366971 | 0.012785627 | 0.125415702 | -0.01554875 |
| Rbmx | 0.006940348 | -0.003567116 | -0.029511463 | 0.013550127 | 0.003860258 |
| Rbp1 | 0.013911583 | 0.007480073 | -0.003555841 | 0.021065336 | 0.020774402 |
| Rbpj | 0.027543178 | -0.062446148 | 0.031453563 | -0.115434783 | 0.016535506 |
| Rcsd1 | -7.53E-07 | 3.21E-04 | -0.003685952 | 8.59E-04 | -1.03E-04 |
| Rec114 | -0.006852748 | 0.009727314 | -0.002742344 | -0.010935564 | -0.002199012 |
| Reep5 | -0.02216472 | 0.013937064 | 0.006580697 | 0.10174785 | -0.020488442 |

| | | | | | |
|---|---|---|---|---|---|
| Renbp | -0.022606464 | 0.031770284 | -0.007548334 | 0.003808536 | 5.96E-04 |
| Rest | 0.080729773 | -0.086419407 | 0.098415799 | -0.111866463 | 0.054600858 |
| Retn | -4.56E-04 | -0.003496705 | -0.00118586 | -0.00266834 | -9.42E-04 |
| Rgcc | -1.23E-04 | 5.40E-05 | -0.001507618 | 0.001148847 | -1.16E-04 |
| Rhob | -0.00271999 | -9.68E-04 | -0.001319756 | 0.021485492 | -0.003596422 |
| Rhox13 | -0.009295119 | 0.009125446 | -0.009092543 | -0.01706062 | -0.002421038 |
| Rhox5 | 0.035659938 | 0.025072045 | -0.002408379 | 7.91E-04 | 0.062431653 |
| Rhox6 | 0.002277232 | 0.010706327 | -2.42E-04 | 0.023093509 | 0.00949264 |
| Rhox9 | -5.76E-04 | 0.011862441 | -3.65E-04 | 0.017572754 | 0.007927308 |
| Rif1 | 0.027114801 | -0.115771151 | -0.009695813 | -0.087037284 | 0.038147449 |
| Rimklb | -4.44E-04 | -0.003153116 | -0.027597494 | 0.002098542 | -5.60E-04 |
| Rnf128 | 7.92E-04 | -0.002675248 | 1.25E-04 | 0.033327502 | -0.001180146 |
| Rnf168 | -0.002734639 | -0.026080327 | 6.74E-04 | -0.011487862 | -0.011861993 |
| Rnf7 | 0.054592704 | -0.005737408 | 0.053069113 | -0.080247834 | 0.057179714 |
| RP23-4H17.3 | 0.010321917 | -0.088433332 | -0.02611706 | -0.126065696 | -0.034539983 |
| Rpl10l | 0.007247167 | 0.001468115 | -0.017599384 | -0.030908162 | 0.007469238 |
| Rpl22l1 | -0.079527423 | 0.020661345 | -0.00827074 | -0.069740476 | -0.080036214 |
| Rpl39l | -0.014857626 | -0.014501862 | -0.021177798 | -0.217748089 | -0.047485167 |
| Rpl41 | -0.015053818 | -0.098884662 | 0.009787067 | -0.023504506 | -0.059752669 |
| Rpp25 | 0.005355047 | -0.001733399 | -0.008759533 | -0.14941101 | 0.010291186 |
| Rps18 | -0.090305688 | 0.090650652 | -0.010222382 | -0.043142608 | -0.058932614 |
| Rps19 | -0.095056839 | 0.027932636 | -0.015859715 | -0.003411023 | -0.084068623 |
| Rps23 | -0.086062758 | 0.042136469 | -0.024822171 | -0.042777868 | -0.068318519 |
| Rps27l | -0.082688969 | -0.043976659 | 0.041606466 | -0.006738416 | -0.103302441 |
| Rps4x | -0.138771111 | 0.16270428 | -0.031896199 | -0.044293326 | -0.08087072 |
| Rrad | -3.59E-05 | -1.57E-04 | -6.87E-04 | 0.002155186 | -1.70E-04 |
| Rrm2 | 0.062996395 | 0.081271004 | 0.004232798 | 0.020020293 | 0.023186878 |
| Rrn3 | 0.05143561 | -0.125552139 | 0.00751245 | -0.033101445 | -0.132459494 |
| Rsph6a | -0.001442344 | 8.94E-04 | -0.001143387 | 7.76E-04 | -9.46E-04 |
| Rsrp1 | -0.014701825 | 0.022407166 | -0.023871414 | 0.004080624 | -0.017995192 |
| Rtkn | -0.032212705 | 0.020219526 | 0.001653975 | -0.008082104 | -0.02648713 |
| S100a1 | -0.001032256 | 6.85E-04 | -0.00141131 | 0.003317158 | -4.08E-04 |
| S100a10 | -0.007515776 | 0.010709035 | -0.016677122 | 0.134839385 | -0.024146248 |
| S100a11 | -0.017622323 | 0.038308948 | -0.051647414 | 0.104227755 | 0.007144318 |
| S100a13 | -0.005313137 | 0.005699655 | -0.004220947 | 0.009745405 | -9.20E-04 |
| S100a3 | -7.28E-05 | -4.79E-05 | -8.84E-04 | 9.61E-04 | -8.22E-05 |
| S100a6 | -0.01750986 | 0.0208029 | 0.009223655 | 0.183407504 | 9.56E-04 |
| Sall3 | -0.006972567 | -0.003315901 | 0.001829552 | -0.002593282 | -0.00583501 |

| | | | | | |
|---|---|---|---|---|---|
| Sall4 | 0.080407128 | -0.141551666 | -0.023251011 | -0.029562528 | -0.075382779 |
| Sat1 | -0.037497097 | 0.002519412 | -0.027491711 | 0.012912893 | -0.038209509 |
| Sbsn | -0.003396512 | 0.001778803 | -4.66E-04 | 0.014858821 | -0.003388966 |
| Scamp1 | -0.001924287 | -9.19E-04 | -0.093333216 | 0.013001173 | -0.002014371 |
| Scd2 | -0.002731334 | -0.035780064 | -4.53E-04 | 0.003550758 | -0.024847525 |
| Scn1b | -0.015183191 | 0.009995915 | 0.001367609 | 0.011313418 | -0.012533951 |
| Sct | -0.00854992 | 0.007564779 | -0.005478534 | 0.018283712 | -0.005161031 |
| Sdc4 | -0.114808771 | 0.056513577 | 0.007923624 | 0.058269913 | -0.150240629 |
| Sdhaf3 | 0.00154874 | -0.007239073 | -0.063608358 | -0.005147507 | -0.001445062 |
| Sec24a | 0.001534855 | -0.017190496 | -0.009297395 | -0.009214654 | -0.008498334 |
| Sema6a | -0.001229824 | -0.001598465 | -5.55E-04 | 0.008687094 | -0.001584807 |
| Sepp1 | -0.002588991 | -6.80E-04 | -0.003069043 | 0.01317601 | -0.002579083 |
| Sep-01 | -0.021865864 | 5.56E-04 | 0.045532499 | -0.230085005 | -0.038977249 |
| Sepw1 | 0.024886366 | -0.090775518 | 0.034786669 | 0.093957501 | -0.019362021 |
| Serpinb9b | 3.07E-04 | -1.60E-04 | 2.08E-04 | 0.009423288 | 1.57E-04 |
| Serpine2 | -0.086982158 | 0.049722233 | -0.007121067 | 0.025940291 | -0.065875181 |
| Serpinh1 | -0.010104198 | 0.002705404 | 0.00320555 | 0.094037236 | -0.014602483 |
| Sfmbt2 | 0.010677992 | -0.011342793 | 0.004524032 | 0.044104618 | 0.009673981 |
| Sfn | 0.016263553 | 0.004312064 | 0.002058849 | 0.21397529 | 0.011694449 |
| Sfrp1 | 0.002607389 | -0.004371941 | 0.008468852 | -0.012738665 | 0.006108912 |
| Sgk1 | 0.030643911 | -0.015206026 | 0.023009411 | -0.040170185 | 0.057098495 |
| Shkbp1 | -3.56E-04 | 0.001220082 | 0.004159807 | 0.01120087 | -2.57E-04 |
| Sin3b | 0.008293403 | 0.061675554 | -0.018001246 | 0.063286411 | 0.040625854 |
| Six1 | 2.67E-04 | -1.55E-04 | -0.001374347 | 0.001895693 | 8.96E-05 |
| Skil | 0.03780045 | -0.116804184 | 0.009948711 | 0.110930305 | -0.099984465 |
| Slamf1 | 0.013278391 | -0.040830895 | -0.014377289 | -0.035838055 | -0.003114401 |
| Slbp | 0.134867037 | 0.01189766 | 0.005888011 | 0.027548431 | 0.049909169 |
| Slc11a1 | -0.007900244 | 0.00174719 | 0.00439865 | -0.026616907 | -0.007437181 |
| Slc16a3 | 0.058979401 | -0.016444153 | 0.013551708 | 0.109715033 | 0.040361413 |
| Slc24a5 | 0.074845948 | -0.167422159 | 0.017423379 | 0.049149133 | -0.186476267 |
| Slc25a20 | -0.005765402 | 0.012164416 | 1.86E-04 | -0.019133076 | 0.003751344 |
| Slc25a31 | -0.001520281 | 0.003237038 | -0.008973264 | 0.001638449 | 4.39E-05 |
| Slc25a4 | 0.055833533 | 0.042460413 | -0.013076396 | 0.204685089 | 0.067176965 |
| Slc25a43 | -9.17E-05 | -2.17E-06 | -8.56E-04 | 8.85E-04 | -2.60E-05 |
| Slc25a5 | 0.04325747 | 0.120312757 | 0.001078678 | -0.064101328 | 0.098645001 |
| Slc26a2 | -0.00489575 | -0.002593123 | -0.002238059 | 0.006411679 | -0.005608339 |
| Slc27a2 | 0.00456625 | -0.007423132 | -0.039019586 | 0.014501676 | 0.004642233 |
| Slc29a1 | 0.029150164 | 0.003859143 | 0.035823688 | -0.092939949 | 0.0379958 |

| | | | | | |
|---|---|---|---|---|---|
| Slc2a1 | 0.050550192 | 0.007672441 | 0.011099045 | 0.117551492 | 0.038046367 |
| Slc2a3 | 0.022019635 | -0.01348105 | -0.01307663 | 0.068712127 | 0.054511664 |
| Slc30a2 | 0.003449103 | 0.001397747 | 0.004264478 | 6.76E-04 | 0.004906282 |
| Slc39a1 | 0.031378041 | -0.063565396 | 0.003622627 | 0.01787054 | -0.078946031 |
| Slc39a8 | 9.74E-04 | -0.002882923 | 1.29E-04 | 0.01229873 | 4.25E-05 |
| Slc40a1 | -5.67E-04 | 0.00185359 | -7.23E-04 | 0.015604466 | 0.001381626 |
| Slc7a3 | 0.033306032 | 0.001392668 | 0.022267516 | -0.035557843 | 0.045667229 |
| Smad7 | -0.005953456 | -0.010465987 | 0.015066023 | 0.015158022 | -0.019193387 |
| Smagp | 0.041124727 | -0.059691403 | 0.046757654 | 0.005657426 | -0.02059663 |
| Smarca5 | 3.02E-04 | -0.08743006 | -0.029486218 | -0.039434529 | -0.01474764 |
| Sms | 0.105968568 | -0.075107269 | 0.033188686 | -0.033721125 | 0.1013138 |
| Smtnl1 | -0.021970016 | 0.012401182 | 0.004190501 | -0.022991177 | -0.01892732 |
| Snai1 | -1.30E-06 | 4.76E-04 | -0.013332754 | 0.003742773 | 3.82E-05 |
| Snhg12 | -0.001269504 | -0.019441422 | -0.005831517 | 7.93E-04 | -0.007852681 |
| Snrpn | 0.086532432 | -0.096974998 | 0.004306377 | 0.090750509 | 0.050789609 |
| Snx20 | -4.35E-04 | 2.98E-04 | -8.36E-04 | 6.28E-04 | -3.26E-04 |
| Socs2 | 0.017930513 | -0.037336786 | 0.036709165 | 0.032684782 | 0.005442333 |
| Socs3 | -0.008710522 | 0.001988631 | -0.016807844 | -0.034478184 | -0.012466812 |
| Sod2 | 0.02827312 | 0.017783941 | 0.038565632 | -0.199847135 | 0.042087259 |
| Soga3 | 0.005238877 | -0.023550209 | 0.006491161 | -0.019517887 | -0.007463357 |
| Sord | -0.02834973 | 0.023346322 | -0.041313672 | -0.004114073 | -0.015661596 |
| Sox11 | 0.005004106 | -0.03077355 | 0.00110728 | 0.053960175 | -0.007388504 |
| Sox15 | -0.013305702 | 0.017353438 | 0.007362039 | -0.028173918 | -0.002028375 |
| Sox2 | -0.032292684 | -0.041332289 | 0.089978823 | -0.14662022 | 0.006670554 |
| Sox21 | -0.003021545 | 0.002847231 | 0.003246959 | -0.007234857 | 0.005735028 |
| Sox4 | -0.012998207 | -0.038595687 | 0.021076517 | 0.078066104 | -0.02832406 |
| Sp110 | 1.62E-04 | -2.43E-04 | -0.055521075 | -2.02E-04 | 9.27E-05 |
| Sp140 | -1.27E-04 | -1.12E-04 | -0.012564749 | 5.50E-04 | -9.54E-05 |
| Sp5 | 0.00173884 | -0.006261732 | 8.16E-04 | 0.016212385 | -5.79E-04 |
| Sparc | -0.011226458 | -0.004374996 | -0.013551364 | 0.069092008 | 0.004118967 |
| Spesp1 | -9.45E-06 | -1.87E-05 | -0.018550628 | 2.52E-04 | 5.46E-06 |
| Spic | -0.001084913 | 0.001690502 | -0.002864992 | -7.47E-04 | 0.001036581 |
| Spink1 | 0.009202317 | -0.026808138 | 0.010753896 | -0.142282087 | -0.00450494 |
| Spp1 | 0.063499584 | 0.081710704 | 0.042511659 | -0.179542182 | 0.148766092 |
| Spry2 | 0.06903804 | -0.166660288 | 0.0138726 | 0.001953989 | -0.191079438 |
| Spry4 | -0.002729379 | -0.052853531 | 0.010859661 | -0.02658631 | -0.027693194 |
| Spty2d1 | -0.023417455 | -0.010474751 | -0.022905467 | -0.046632767 | -0.016797284 |
| Sqstm1 | -0.019603718 | 0.032103436 | -0.016705487 | 0.042550029 | -0.004273325 |

| | | | | | |
|---|---|---|---|---|---|
| Srgn | -5.41E-05 | 1.40E-05 | -8.16E-04 | 4.34E-04 | -5.99E-05 |
| Stk32c | -2.60E-05 | 6.87E-06 | -9.31E-04 | 6.82E-04 | -4.88E-05 |
| Stk35 | 0.002181807 | -0.001218567 | -0.02126966 | -0.004155521 | 7.39E-05 |
| Stmn1 | -0.006790149 | 0.094814197 | 0.031683501 | -0.275261612 | 0.045357597 |
| Stmn2 | 0.01748831 | 0.011893068 | 0.040464023 | -0.171369153 | -0.006780733 |
| Stox1 | -3.05E-04 | 1.26E-04 | -0.001621905 | 5.60E-04 | -2.82E-04 |
| Stra8 | 0.00163326 | 0.007369675 | -0.003663478 | -0.020353357 | 0.004136973 |
| Stx3 | -0.022925106 | -0.001614687 | 0.012644452 | 0.09261637 | -0.021189014 |
| Sycp3 | -0.004424388 | 0.010675615 | -0.004789932 | 0.008353822 | 0.003633398 |
| Synm | -0.006236138 | 0.003304678 | -0.0011922 | 0.00312909 | -0.004560674 |
| Taf7l | -0.003042761 | 0.002052299 | -0.004256754 | 0.003927276 | -0.001225465 |
| Tagln | 0.017444811 | -0.027355756 | 0.006003767 | 0.181692801 | 0.003073052 |
| Tagln2 | -0.007055003 | 0.034045821 | 0.008082235 | 0.125109566 | -0.015343894 |
| Tapbpl | 0.006351503 | -0.0215806 | -0.010955962 | -0.014133389 | -0.003086879 |
| Tax1bp3 | 0.003950829 | -0.006255817 | -0.010966673 | 0.09347374 | -0.006094856 |
| Tbc1d12 | -0.001322681 | 6.92E-04 | -0.022061077 | 0.003335758 | -6.88E-04 |
| Tbc1d24 | -0.002585957 | -0.007905513 | 6.98E-04 | -0.008820908 | -0.004767308 |
| Tbx3 | -0.04554128 | 0.035932658 | -0.006373377 | -0.135304745 | -0.0142048 |
| Tc2n | -7.50E-05 | -4.14E-04 | -0.002499406 | 0.001116112 | -2.00E-04 |
| Tcea3 | 0.040390526 | -0.067517276 | 0.057219849 | -0.194945333 | -0.003706459 |
| Tcf15 | -0.016374355 | -0.133389215 | 0.08975039 | -0.002428881 | -0.138806691 |
| Tchh | -1.03E-04 | -5.65E-05 | -0.001647405 | 0.001054748 | -1.08E-04 |
| Tcl1 | 0.004666572 | 0.017234011 | 0.019406218 | -0.123468508 | 0.020166421 |
| Tctn1 | -6.22E-04 | -6.81E-04 | -0.00242338 | 7.03E-04 | -0.001254814 |
| Tdgf1 | 0.035548115 | -0.060323465 | 0.051029405 | -0.069655398 | 0.024517603 |
| Tdh | -0.019114841 | 0.0254043 | 0.035266166 | -0.136562069 | 0.009687759 |
| Terf2ip | 0.00329597 | 0.001253088 | 9.41E-05 | 9.89E-04 | -4.41E-04 |
| Tet1 | -0.011376711 | -0.126617407 | 0.057614585 | -0.179859118 | -0.019799202 |
| Tex101 | -1.96E-04 | 3.81E-04 | -0.002446315 | 6.95E-04 | 7.23E-05 |
| Tex19.1 | 0.030302796 | 0.053691203 | -0.072890754 | -0.050399334 | 0.044186459 |
| Tfap2c | -0.003659124 | 0.021629114 | 0.012726465 | 0.009844342 | 0.019641186 |
| Tfpi | 0.005837204 | 0.014491226 | 0.009665238 | -0.020821894 | 0.014611721 |
| Tgfb1i1 | -8.26E-04 | -5.17E-04 | -0.002733752 | 0.012193504 | -0.001283558 |
| Thbs1 | 1.95E-04 | -0.002092082 | -5.50E-04 | 0.014570099 | -0.001111871 |
| Thnsl2 | -7.44E-04 | 0.00273594 | -0.021792891 | -3.16E-04 | 1.30E-04 |
| Ticrr | -0.005335598 | -0.002250029 | -0.030422961 | -0.048645393 | 0.008064349 |
| Timp2 | 0.001110403 | -2.86E-04 | -0.002828483 | 0.017401259 | -7.60E-04 |
| Tinagl1 | -3.47E-04 | -9.28E-04 | 2.44E-05 | 0.03028404 | -0.00166679 |

| | | | | | |
|---|---|---|---|---|---|
| Tmcc1 | -0.002629923 | -0.008119509 | -0.002109172 | 0.004069029 | -0.006689455 |
| Tmem191c | -0.004632722 | 0.007886984 | -0.017643003 | -0.01986261 | 0.005076412 |
| Tmem245 | 0.061914422 | -0.142560247 | -0.024572116 | -0.14440798 | -0.021842091 |
| Tmem37 | -0.002575682 | 0.0043391 | 0.004484148 | 0.013929301 | -0.001752279 |
| Tmem40 | -0.005343301 | -0.002262128 | 0.010548602 | -0.031415019 | -0.00415688 |
| Tmem59 | -0.001840633 | 0.050978679 | -0.02265903 | 0.084140141 | 0.028300742 |
| Tmem92 | 5.34E-04 | -1.67E-05 | -0.059765857 | 3.38E-05 | 9.59E-06 |
| Tmsb4x | -0.011841952 | -0.087936215 | 0.037345172 | -0.055501702 | -0.054210212 |
| Tnfaip6 | -2.24E-04 | 1.17E-04 | -9.50E-04 | 7.04E-04 | -2.08E-04 |
| Tnfrsf12a | 0.002458745 | -0.009743549 | -0.00206313 | 0.086160446 | -0.006528297 |
| Top2a | -0.035564008 | -0.15278487 | 0.023615703 | -0.050582968 | 0.008013693 |
| Tpbg | -8.89E-04 | -0.001327996 | -0.001964892 | 0.02399449 | -0.001321577 |
| Tpi1 | 0.104392092 | 0.030268297 | 0.056830361 | 0.094655648 | 0.09177497 |
| Tpm1 | 0.074709894 | -0.074961246 | 0.034140516 | 0.305329183 | 0.062319308 |
| Tpm4 | 0.016981044 | -0.023871927 | 0.01636692 | 0.159100565 | 0.004022734 |
| Tppp3 | -2.33E-04 | 0.005032102 | -0.00287463 | -0.007627871 | 0.004574053 |
| Trap1a | 0.0251173 | 0.043527441 | 0.011703086 | -0.072774944 | 0.042923267 |
| Trh | 0.017331466 | 1.11E-05 | 0.041924522 | -0.045260002 | 0.032257863 |
| Trib3 | -0.01326889 | 0.005559874 | 0.007102837 | 0.011460562 | 4.13E-04 |
| Trim47 | -6.27E-04 | 6.40E-04 | -9.08E-04 | 0.001807376 | -2.92E-04 |
| Trp53i11 | -0.018744991 | 0.004303259 | 0.006958641 | 0.004838434 | -0.018849636 |
| Trp53inp1 | -0.104020595 | 0.055888927 | -0.011292357 | 0.003549626 | -0.087639199 |
| Tsc22d1 | -0.031893044 | 0.007198355 | -0.002233002 | 0.006854731 | -0.034580682 |
| Tsen2 | 0.016824822 | 0.001894969 | -0.076650506 | 0.002420435 | 0.016838644 |
| Ttc39c | -1.88E-04 | -6.53E-05 | -0.001462863 | 7.57E-04 | -2.59E-04 |
| Ttn | 0.00415939 | -0.041762918 | 0.009034515 | -0.029935234 | 0.003746578 |
| Tuba1a | -0.004364987 | 0.047298235 | -0.005127153 | 0.188470036 | 0.038465813 |
| Tuba3a | 0.0025798 | 0.025986628 | -0.008402449 | -0.044935185 | 0.011337176 |
| Tubb2a | 0.008988741 | 0.01421165 | -0.006732093 | 8.09E-04 | 0.015439416 |
| Tubb3 | 0.043099844 | 0.055410009 | 0.001802599 | -0.068141474 | 0.073497107 |
| Tubb4b | 0.051455516 | 0.102900773 | -0.035019982 | 0.00457642 | 0.094496539 |
| Tubb5 | 0.080750371 | 0.113805303 | 0.001125909 | 0.061903665 | 0.118368445 |
| Tubb6 | -0.013510095 | 0.04087267 | 0.006376631 | 0.218677916 | -0.026298526 |
| Txnip | -0.022608507 | -0.008714807 | 0.004532519 | 0.016844348 | -0.021175459 |
| Uap1l1 | -0.024313752 | 0.022253278 | -0.004798258 | 0.018057994 | -0.006834176 |
| Ubald2 | -0.070053764 | 0.050123561 | 0.070642731 | -0.092903931 | -0.025682894 |
| Ube2c | 0.001227941 | 0.01489761 | -0.011420286 | -0.06547955 | 0.077675371 |
| Ung | 0.084521108 | 0.024556238 | 0.040033 | -0.042632856 | 0.024663399 |

| | | | | | |
|---|---|---|---|---|---|
| Upp1 | 0.045687251 | 0.109770207 | 0.022351227 | -0.21130694 | 0.094598683 |
| Uqcrb | 0.010727709 | 0.033637388 | 0.001193446 | -0.02526016 | 0.037841921 |
| Usp13 | -1.74E-04 | -5.02E-05 | -0.001423277 | 0.001294684 | -1.65E-04 |
| Usp17la | 5.27E-04 | 1.28E-04 | -0.085861938 | -4.57E-05 | -9.95E-05 |
| Usp17lc | 1.19E-04 | -2.36E-04 | -0.039888132 | -3.40E-05 | 6.13E-06 |
| Usp19 | 0.003477113 | -0.051984341 | -0.015792072 | -0.052333743 | -0.019190364 |
| Usp26 | 2.35E-04 | -3.30E-04 | -0.01526298 | 0.001379098 | 1.97E-04 |
| Utf1 | 0.125506859 | -0.085249583 | 0.137647299 | -0.039441996 | 0.051482642 |
| Vamp8 | -0.005035676 | 6.99E-04 | 0.003879445 | 0.064910437 | -0.011927243 |
| Vgll3 | -5.66E-04 | -0.001046278 | -2.39E-04 | 0.016802684 | -9.84E-04 |
| Vim | -0.006004467 | 0.031547737 | 0.001697801 | -0.043028038 | -0.013009008 |
| Wbp5 | 0.021773584 | 0.010544356 | 0.021040523 | 0.114374668 | 0.036512312 |
| Wfdc2 | 0.017999323 | 0.010202317 | 0.002728327 | 0.030126048 | 0.026044562 |
| Wls | -0.002000118 | 5.56E-04 | -0.005939198 | 0.009773865 | -0.001372056 |
| Wnt3 | -2.58E-04 | -5.72E-05 | -0.001056892 | 0.003331351 | -3.77E-04 |
| Wnt4 | -2.34E-04 | -4.23E-04 | -3.22E-04 | 0.006292003 | -6.88E-04 |
| Wnt6 | -0.009638128 | 0.004688437 | 0.001359393 | -0.001158489 | -0.01205895 |
| Wnt7b | -2.66E-04 | 9.11E-05 | -5.32E-04 | 0.007474071 | -2.05E-04 |
| Wtap | 0.064959142 | -0.137986609 | 0.006353335 | -0.018111977 | -0.150579208 |
| Xist | -0.002635861 | -0.001732221 | -0.001707079 | 0.002448734 | -0.008506476 |
| Xlr3a | -7.30E-05 | 9.88E-05 | -0.00119254 | 6.24E-04 | 3.43E-05 |
| Ywhaq | 0.100143144 | 0.031235185 | -0.006362319 | 0.019520464 | 0.089766266 |
| Zbtb10 | -0.00139341 | 8.02E-04 | -0.004099092 | 0.007061992 | 2.66E-04 |
| Zc3hav1 | 0.008614949 | -0.006964434 | -0.008821888 | -0.001053405 | -0.011471245 |
| Zfos1 | -0.02421742 | -0.029596767 | 0.004043892 | 0.01624874 | -0.033703985 |
| Zfp287 | -2.17E-04 | -3.58E-04 | -0.001487744 | 0.00109094 | -2.13E-04 |
| Zfp296 | 0.016364233 | 0.025446963 | -0.05429264 | -0.068131632 | 0.040030117 |
| Zfp316 | -1.44E-04 | -0.001385645 | -0.001937904 | -1.55E-04 | -7.10E-04 |
| Zfp326 | 0.003469998 | -0.017138152 | 0.012520657 | 0.035213075 | 1.51E-05 |
| Zfp42 | 0.021628098 | -0.00124428 | 0.108338397 | -0.501610332 | 0.088986416 |
| Zfp428 | 0.093837115 | -0.108783132 | 0.049686184 | -0.075297815 | 0.018790598 |
| Zfp516 | -0.004558494 | -0.003002231 | -0.050578507 | -0.00647421 | -0.001952521 |
| Zfp560 | -1.05E-04 | -0.017911743 | -0.137954837 | 0.005416308 | -0.008261478 |
| Zfp57 | 0.009088107 | -0.011049814 | 0.028582625 | -0.150061246 | 0.022108253 |
| Zfp706 | 0.052316871 | -0.079185658 | 0.070999492 | 0.05294064 | -0.011841548 |
| Zfp809 | -0.00238174 | -7.82E-04 | -0.048392675 | -0.001829533 | -0.002788604 |
| Zic2 | 0.003889626 | -0.006410234 | 0.00477635 | 0.001573961 | 0.002224383 |
| Zic5 | 0.011087826 | -0.020305948 | 0.009983686 | 0.008020038 | 0.008656391 |

| | | | | | |
|---|---|---|---|---|---|
| Zscan4b | 1.14E-05 | -4.60E-05 | -0.002376454 | 4.10E-04 | -1.69E-05 |
| Zscan4c | 0.001862291 | 0.001310167 | -0.248795914 | -9.17E-04 | -8.11E-04 |
| Zscan4d | 0.002096548 | 5.82E-04 | -0.214888328 | -8.93E-04 | -7.41E-04 |
| Zscan4f | 8.20E-04 | 0.002450292 | -0.274776455 | -0.001061091 | -0.001421238 |
| Zyx | 0.008417029 | -0.024142531 | 0.004989297 | 0.109270258 | -0.006464848 |

# Appendix H Gene ontology analysis for MOFA factors 1-5

Gene Ontology (GO) terms enriched for the top 50 gene loadings for MOFA factor 1, using 965 highly variable genes as a background:

| **Process** | | | | | |
|---|---|---|---|---|---|
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0043604 | amide biosynthetic process | 1.72E-08 | 1.27E-04 | 10.33 (900,16,49,9) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| | | | | | Ass1 - argininosuccinate synthetase 1 |
| GO:0006412 | translation | 1.83E-08 | 6.75E-05 | 12.24 (900,12,49,8) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| GO:0043043 | peptide biosynthetic process | 1.83E-08 | 4.50E-05 | 12.24 (900,12,49,8) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| GO:0072332 | intrinsic apoptotic signaling pathway by p53 class mediator | 4.96E-07 | 9.17E-04 | 13.78 (900,8,49,6) | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Ddit4 - dna-damage-inducible transcript 4 |
| | | | | | Perp - perp, tp53 apoptosis effector |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| GO:0006518 | peptide metabolic process | 9.17E-07 | 1.36E-03 | 7.19 (900,23,49,9) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Gpx3 - glutathione peroxidase 3 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| GO:0072331 | signal transduction by p53 class mediator | 1.24E-06 | 1.52E-03 | 9.89 (900,13,49,7) | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Ddit4 - dna-damage-inducible transcript 4 |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Perp - perp, tp53 apoptosis effector |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| GO:0043603 | cellular amide metabolic process | 4.41E-06 | 4.66E-03 | 5.40 (900,34,49,10) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Gpx3 - glutathione peroxidase 3 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Ass1 - argininosuccinate synthetase 1 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| GO:0042771 | intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator | 7.57E-06 | 7.00E-03 | 13.12 (900,7,49,5) | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Ddit4 - dna-damage-inducible transcript 4 |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| GO:0045930 | negative regulation of mitotic cell cycle | 1.64E-05 | 1.35E-02 | 6.12 (900,24,49,8) | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Btg2 - b cell translocation gene 2, anti-proliferative |
| | | | | | Btg1 - b cell translocation gene 1, anti-proliferative |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Plk2 - polo-like kinase 2 |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Ccng1 - cyclin g1 |
| GO:0045786 | negative regulation of cell cycle | 8.34E-05 | 6.17E-02 | 3.99 (900,46,49,10) | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Btg2 - b cell translocation gene 2, anti-proliferative |
| | | | | | Btg1 - b cell translocation gene 1, anti-proliferative |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Hsp90ab1 - heat shock protein 90 alpha (cytosolic), class b member 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Plk2 - polo-like kinase 2 |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Ccng1 - cyclin g1 |
| | | | | | Trp53inp1 - transformation related protein 53 inducible nuclear protein 1 |
| GO:0044271 | cellular nitrogen compound biosynthetic process | 1.51E-04 | 1.02E-01 | 3.19 (900,69,49,12) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Pmm1 - phosphomannomutase 1 |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Hsp90ab1 - heat shock protein 90 alpha (cytosolic), class b member 1 |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| | | | | | Ass1 - argininosuccinate synthetase 1 |
| GO:0007346 | | 1.78E-04 | 1.10E-01 | 3.67 (900,50,49,10) | Psrc1 - proline/serine-rich coiled-coil 1 |

| | | | | | |
|---|---|---|---|---|---|
| | regulation of mitotic cell cycle | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Btg2 - b cell translocation gene 2, anti-proliferative |
| | | | | | Btg1 - b cell translocation gene 1, anti-proliferative |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Plk2 - polo-like kinase 2 |
| | | | | | Cenpf - centromere protein f |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Ccng1 - cyclin g1 |
| GO:0034645 | cellular macromolecule biosynthetic process | 2.52E-04 | 1.43E-01 | 3.53 (900,52,49,10) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Hsp90ab1 - heat shock protein 90 alpha (cytosolic), class b member 1 |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| GO:0006974 | cellular response to | 3.31E-04 | 1.75E-01 | 3.76 (900,44,49,9) | Btg2 - b cell translocation gene 2, anti-proliferative |

| | | | | | |
|---|---|---|---|---|---|
| | DNA damage stimulus | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Ddit4 - dna-damage-inducible transcript 4 |
| | | | | | 1700007K13Rik - riken cdna 1700007k13 gene |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Ccng1 - cyclin g1 |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| GO:1901566 | organonitrogen compound biosynthetic process | 3.31E-04 | 1.63E-01 | 3.76 (900,44,49,9) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Ass1 - argininosuccinate synthetase 1 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| GO:0008630 | intrinsic apoptotic signaling pathway in response to DNA damage | 3.61E-04 | 1.67E-01 | 7.06 (900,13,49,5) | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Ddit4 - dna-damage-inducible transcript 4 |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| GO:1901991 | negative regulation of mitotic cell cycle phase transition | 3.61E-04 | 1.57E-01 | 7.06 (900,13,49,5) | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Ccng1 - cyclin g1 |
| GO:0052548 | regulation of endopeptidase activity | 4.07E-04 | 1.67E-01 | 4.08 (900,36,49,8) | Fam162a - family with sequence similarity 162, member a |
| | | | | | Serpine2 - serine (or cysteine) peptidase inhibitor, clade e, member 2 |
| | | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Ctsd - cathepsin d |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Perp - perp, tp53 apoptosis effector |
| GO:2000116 | regulation of cysteine-type endopeptidase activity | 4.53E-04 | 1.76E-01 | 4.59 (900,28,49,7) | Fam162a - family with sequence similarity 162, member a |
| | | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Ctsd - cathepsin d |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Perp - perp, tp53 apoptosis effector |
| GO:1902807 | negative regulation of cell cycle G1/S phase transition | 4.64E-04 | 1.72E-01 | 9.18 (900,8,49,4) | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| GO:2000134 | negative regulation of G1/S transition of mitotic cell cycle | 4.64E-04 | 1.63E-01 | 9.18 (900,8,49,4) | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| GO:1901988 | negative regulation of cell cycle phase transition | 5.39E-04 | 1.81E-01 | 6.56 (900,14,49,5) | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Ccng1 - cyclin g1 |
| GO:0010948 | negative regulation of cell cycle process | 7.33E-04 | 2.36E-01 | 5.01 (900,22,49,6) | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Hsp90ab1 - heat shock protein 90 alpha (cytosolic), class b member 1 |
| | | | | | Rps27l - ribosomal protein s27-like |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Ccng1 - cyclin g1 |
| GO:0030162 | regulation of proteolysis | 8.60E-04 | 2.65E-01 | 3.06 (900,60,49,10) | Fam162a - family with sequence similarity 162, member a |
| | | | | | Serpine2 - serine (or cysteine) peptidase inhibitor, clade e, member 2 |
| | | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Hsp90ab1 - heat shock protein 90 alpha (cytosolic), class b member 1 |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Ctsd - cathepsin d |
| | | | | | Plk2 - polo-like kinase 2 |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Perp - perp, tp53 apoptosis effector |
| GO:0009059 | macromolecule biosynthetic process | 8.60E-04 | 2.54E-01 | 3.06 (900,60,49,10) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Hsp90ab1 - heat shock protein 90 alpha (cytosolic), class b member 1 |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Rps18 - ribosomal protein s18 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| GO:0044267 | cellular protein metabolic process | 9.10E-04 | 2.59E-01 | 2.40 (900,107,49,14) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Pmm1 - phosphomannomutase 1 |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Eef1a1 - eukaryotic translation elongation factor 1 alpha 1 |
| | | | | | Plk2 - polo-like kinase 2 |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Eef2 - eukaryotic translation elongation factor 2 |
| | | | | | Ptp4a3 - protein tyrosine phosphatase 4a3 |
| | | | | | Btg2 - b cell translocation gene 2, anti-proliferative |
| | | | | | Btg1 - b cell translocation gene 1, anti-proliferative |
| | | | | | Rps19 - ribosomal protein s19 |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Rps18 - ribosomal protein s18 |
| GO:0043085 | positive regulation of catalytic activity | 9.41E-04 | 2.58E-01 | 2.51 (900,95,49,13) | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Chchd10 - coiled-coil-helix-coiled-coil-helix domain containing 10 |
| | | | | | Fam162a - family with sequence |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | similarity 162, member a |
| | | | | | Psrc1 - proline/serine-rich coiled-coil 1 |
| | | | | | Btg1 - b cell translocation gene 1, anti-proliferative |
| | | | | | Cd81 - cd81 antigen |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Hsp90ab1 - heat shock protein 90 alpha (cytosolic), class b member 1 |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Ctsd - cathepsin d |
| | | | | | Perp - perp, tp53 apoptosis effector |
| | | | | | Sdc4 - syndecan 4 |
| GO:0097193 | intrinsic apoptotic signaling pathway | 9.51E-04 | 2.51E-01 | 4.79 (900,23,49,6) | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Ddit4 - dna-damage-inducible transcript 4 |
| | | | | | Perp - perp, tp53 apoptosis effector |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| **Function** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0003735 | structural constituent of ribosome | 1.43E-06 | 1.97E-03 | 12.24 (900,9,49,6) | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Rps19 - ribosomal protein s19<br>Rps18 - ribosomal protein s18 |
| GO:0019843 | rRNA binding | 7.80E-06 | 5.38E-03 | 18.37 (900,4,49,4) | Rps4x - ribosomal protein s4, x-linked<br>Mdm2 - transformed mouse 3t3 cell double minute 2<br>Rps18 - ribosomal protein s18<br>Eef2 - eukaryotic translation elongation factor 2 |
| **Component** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0005840 | ribosome | 2.58E-7 | 2.2E-4 | 11.69 (900,11,49,7) | Rps4x - ribosomal protein s4, x-linked<br>Rpl22l1 - ribosomal protein l22 like 1<br>Rps27l - ribosomal protein s27-like<br>Rps23 - ribosomal protein s23<br>Rps19 - ribosomal protein s19<br>Rps18 - ribosomal protein s18<br>Eef2 - eukaryotic translation elongation factor 2 |
| GO:0022627 | cytosolic small ribosomal subunit | 3.92E-7 | 1.67E-4 | 18.37 (900,5,49,5) | Rps4x - ribosomal protein s4, x-linked<br>Rps27l - ribosomal protein s27-like<br>Rps23 - ribosomal protein s23<br>Rps19 - ribosomal protein s19<br>Rps18 - ribosomal protein s18 |
| GO:0015935 | small ribosomal subunit | 3.92E-7 | 1.11E-4 | 18.37 (900,5,49,5) | Rps4x - ribosomal protein s4, x-linked<br>Rps27l - ribosomal protein s27-like<br>Rps23 - ribosomal |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | protein s23<br>Rps19 - ribosomal protein s19<br>Rps18 - ribosomal protein s18 |
| GO:0044391 | ribosomal subunit | 1.43E-6 | 3.03E-4 | 12.24 (900,9,49,6) | Rps4x - ribosomal protein s4, x-linked<br>Rpl22l1 - ribosomal protein l22 like 1<br>Rps27l - ribosomal protein s27-like<br>Rps23 - ribosomal protein s23<br>Rps19 - ribosomal protein s19<br>Rps18 - ribosomal protein s18 |
| GO:0044445 | cytosolic part | 2.46E-5 | 4.19E-3 | 8.48 (900,13,49,6) | Rps4x - ribosomal protein s4, x-linked<br>Rpl22l1 - ribosomal protein l22 like 1<br>Rps27l - ribosomal protein s27-like<br>Rps23 - ribosomal protein s23<br>Rps19 - ribosomal protein s19<br>Rps18 - ribosomal protein s18 |

Gene Ontology (GO) terms enriched fpr the top 50 gene loadings for MOFA factor 2, using 965 highly variable genes as a background:

| **Process** | | | | | |
|---|---|---|---|---|---|
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:2001251 | negative regulation of chromosome organization | 2.81E-06 | 2.08E-02 | 9.26 (900,17,40,7) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Top2a - topoisomerase (dna) ii alpha |
| GO:0033044 | regulation of chromosome organization | 1.20E-05 | 4.45E-02 | 6.43 (900,28,40,8) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Top2a - topoisomerase (dna) ii alpha |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |
| GO:1905268 | negative regulation of chromatin organization | 1.49E-05 | 3.68E-02 | 12.50 (900,9,40,5) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| GO:0032259 | methylation | 3.04E-05 | 5.62E-02 | 8.44 (900,16,40,6) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Comt - catechol-o-methyltransferase |
| | | | | | Wtap - wilms' tumour 1-associating protein |
| | | | | | Rbm15 - rna binding motif protein 15 |
| | | | | | Meg3 - maternally expressed 3 |
| GO:0040029 | regulation of gene expression, epigenetic | 6.59E-05 | 9.75E-02 | 7.50 (900,18,40,6) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Rbm15 - rna binding motif protein 15 |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |
| GO:0010639 | negative regulation of organelle organization | 1.11E-04 | 1.37E-01 | 4.86 (900,37,40,8) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Top2a - topoisomerase (dna) ii alpha |
| | | | | | Tmsb4x - thymosin, beta 4, x chromosome |
| GO:1902275 | regulation of chromatin organization | 1.29E-04 | 1.36E-01 | 6.75 (900,20,40,6) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |
| GO:0043414 | macromolecule methylation | 2.01E-04 | 1.86E-01 | 8.04 (900,14,40,5) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Wtap - wilms' tumour 1-associating protein |
| | | | | | Rbm15 - rna binding motif protein 15 |
| | | | | | Meg3 - maternally expressed 3 |
| GO:0031062 | positive regulation of | 2.07E-04 | 1.70E-01 | 11.25 (900,8,40,4) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |

| | | | | | |
|---|---|---|---|---|---|
| | histone methylation | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |
| GO:0000018 | regulation of DNA recombination | 2.07E-04 | 1.53E-01 | 11.25 (900,8,40,4) | |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |
| GO:0051053 | negative regulation of DNA metabolic process | 2.92E-04 | 1.96E-01 | 7.50 (900,15,40,5) | |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |
| GO:0051570 | regulation of histone H3-K9 methylation | 3.16E-04 | 1.95E-01 | 16.88 (900,4,40,3) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |
| GO:0031058 | positive regulation of histone modification | 5.81E-04 | 3.31E-01 | 9.00 (900,10,40,4) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |
| GO:1905269 | positive regulation of chromatin organization | 5.81E-04 | 3.07E-01 | 9.00 (900,10,40,4) | |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |

| | | | | | |
|---|---|---|---|---|---|
| GO:0060969 | negative regulation of gene silencing | 7.66E-04 | 3.78E-01 | 13.50 (900,5,40,3) | Hist1h1e - histone cluster 1, h1e<br>Tet1 - tet methylcytosine dioxygenase 1<br>Hist1h1b - histone cluster 1, h1b |
| GO:0031935 | regulation of chromatin silencing | 7.66E-04 | 3.54E-01 | 13.50 (900,5,40,3) | Hist1h1e - histone cluster 1, h1e<br>Tet1 - tet methylcytosine dioxygenase 1<br>Hist1h1b - histone cluster 1, h1b |
| GO:0031936 | negative regulation of chromatin silencing | 7.66E-04 | 3.33E-01 | 13.50 (900,5,40,3) | Hist1h1e - histone cluster 1, h1e<br>Tet1 - tet methylcytosine dioxygenase 1<br>Hist1h1b - histone cluster 1, h1b |
| GO:0006304 | DNA modification | 8.84E-04 | 3.63E-01 | 8.18 (900,11,40,4) | Dnmt3b - dna methyltransferase 3b<br>Parp1 - poly (adp-ribose) polymerase family, member 1<br>Tet1 - tet methylcytosine dioxygenase 1<br>Meg3 - maternally expressed 3 |
| **Function** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0003676 | nucleic acid binding | 4.75E-05 | 6.55E-02 | 2.31 (900,195,40,20) | Sall4 - sal-like 4 (drosophila)<br>Parp1 - poly (adp-ribose) polymerase family, member 1<br>Nip7 - nuclear import 7 homolog (s. cerevisiae)<br>Myef2 - myelin basic protein expression factor 2, repressor<br>Skil - ski-like<br>Ncoa3 - nuclear receptor coactivator 3<br>Dnajc21 - dnaj (hsp40) homolog, subfamily c, member 21<br>Jarid2 - jumonji, at rich interactive domain 2<br>Rrn3 - rrn3 rna polymerase i transcription factor homolog (yeast) |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Rbm15 - rna binding motif protein 15 |
| | | | | | L1td1 - line-1 type transposase domain containing 1 |
| | | | | | Tcf15 - transcription factor 15 |
| | | | | | Cbx3 - chromobox 3 |
| | | | | | Zfp428 - zinc finger protein 428 |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Snrpn - small nuclear ribonucleoprotein n |
| | | | | | Top2a - topoisomerase (dna) ii alpha |
| GO:1901363 | heterocyclic compound binding | 3.90E-04 | 2.69E-01 | 1.84 (900,281,40,23) | |
| | | | | | Sall4 - sal-like 4 (drosophila) |
| | | | | | Pfas - phosphoribosylformylglycinamidine synthase (fgar amidotransferase) |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Nip7 - nuclear import 7 homolog (s. cerevisiae) |
| | | | | | Skil - ski-like |
| | | | | | Plk4 - polo-like kinase 4 |
| | | | | | Myef2 - myelin basic protein expression factor 2, repressor |
| | | | | | Ncoa3 - nuclear receptor coactivator 3 |
| | | | | | Dnajc21 - dnaj (hsp40) homolog, subfamily c, member 21 |
| | | | | | Rrn3 - rrn3 rna polymerase i transcription factor homolog (yeast) |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Rbm15 - rna binding motif protein 15 |
| | | | | | L1td1 - line-1 type transposase domain containing 1 |
| | | | | | Tcf15 - transcription factor 15 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Actb - actin, beta |
| | | | | | Cbx3 - chromobox 3 |
| | | | | | Zfp428 - zinc finger protein 428 |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Snrpn - small nuclear ribonucleoprotein n |
| | | | | | Top2a - topoisomerase (dna) ii alpha |
| GO:0097159 | organic cyclic compound binding | 5.54E-04 | 2.54E-01 | 1.80 (900,287,40,23) | |
| | | | | | Sall4 - sal-like 4 (drosophila) |
| | | | | | Pfas - phosphoribosylformylglycinamidine synthase (fgar amidotransferase) |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Nip7 - nuclear import 7 homolog (s. cerevisiae) |
| | | | | | Skil - ski-like |
| | | | | | Plk4 - polo-like kinase 4 |
| | | | | | Myef2 - myelin basic protein expression factor 2, repressor |
| | | | | | Ncoa3 - nuclear receptor coactivator 3 |
| | | | | | Dnajc21 - dnaj (hsp40) homolog, subfamily c, member 21 |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Rrn3 - rrn3 rna polymerase i transcription factor homolog (yeast) |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Rbm15 - rna binding motif protein 15 |
| | | | | | L1td1 - line-1 type transposase domain containing 1 |
| | | | | | Tcf15 - transcription factor 15 |
| | | | | | Actb - actin, beta |
| | | | | | Cbx3 - chromobox 3 |
| | | | | | Zfp428 - zinc finger protein 428 |
| | | | | | Dnmt3b - dna methyltransferase 3b |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Snrpn - small nuclear ribonucleoprotein n |
| | | | | | Top2a - topoisomerase (dna) ii alpha |
| **Component** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0044428 | nuclear part | 2.61E-04 | 2.22E-01 | 1.94 (900,255,40,22) | |
| | | | | | Sall4 - sal-like 4 (drosophila) |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Nip7 - nuclear import 7 homolog (s. cerevisiae) |
| | | | | | Plk4 - polo-like kinase 4 |
| | | | | | Myef2 - myelin basic protein expression factor 2, repressor |
| | | | | | Skil - ski-like |
| | | | | | Ncoa3 - nuclear receptor coactivator 3 |
| | | | | | Fam60a - family with sequence similarity 60, member a |
| | | | | | Jarid2 - jumonji, at rich interactive domain 2 |
| | | | | | Rrn3 - rrn3 rna polymerase i transcription factor homolog (yeast) |
| | | | | | Rbm15 - rna binding motif protein 15 |
| | | | | | Tcf15 - transcription factor 15 |
| | | | | | Actb - actin, beta |
| | | | | | Cbx3 - chromobox 3 |
| | | | | | Dnmt3b - dna methyltransferase 3b |
| | | | | | Hist1h1e - histone cluster 1, h1e |
| | | | | | Wtap - wilms' tumour 1-associating protein |
| | | | | | Hist1h1b - histone cluster 1, h1b |
| | | | | | Snrpn - small nuclear ribonucleoprotein n |
| | | | | | Malat1 - metastasis associated lung adenocarcinoma transcript 1 (non-coding rna) |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Top2a - topoisomerase (dna) ii alpha |
| | | | | | Rif1 - rap1 interacting factor 1 homolog (yeast) |

Gene Ontology (GO) terms enriched fpr the top 50 gene loadings for MOFA factor 3, using 965 highly variable genes as a background:

| **Process:** | None | | | | |
|---|---|---|---|---|---|
| **Function:** | None | | | | |
| **Component:** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0000781 | chromosome, telomeric region | 5.25E-04 | 4.47E-01 | 9.09 (901,9,44,4) | |
| | | | | | Zscan4c - zinc finger and scan domain containing 4c |
| | | | | | Ercc4 - excision repair cross-complementing rodent repair deficiency, complementation group 4 |
| | | | | | Zscan4d - zinc finger and scan domain containing 4d |
| | | | | | Zscan4f - zinc finger and scan domain containing 4f |

Gene Ontology (GO) terms enriched for the top 50 gene loadings for MOFA factor 4, using 965 highly variable genes as a background:

| **Process** | | | | | |
|---|---|---|---|---|---|
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:1902175 | regulation of oxidative stress-induced intrinsic apoptotic | 8.80E-05 | 6.51E-01 | 21.95 (900,3,41,3) | |
| | | | | | Hspb1 - heat shock protein 1 |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |

| | | | | | |
|---|---|---|---|---|---|
| | signaling pathway | | | | Sod2 - superoxide dismutase 2, mitochondrial |
| GO:0071498 | cellular response to fluid shear stress | 3.41E-04 | 1.00E+00 | 16.46 (900,4,41,3) | <br>Klf4 - kruppel-like factor 4 (gut)<br>Klf2 - kruppel-like factor 2 (lung)<br>Spp1 - secreted phosphoprotein 1 |
| GO:1990823 | response to leukemia inhibitory factor | 3.48E-04 | 8.59E-01 | 4.80 (900,32,41,7) | <br>Klf4 - kruppel-like factor 4 (gut)<br>Enah - enabled homolog (drosophila)<br>Zfp42 - zinc finger protein 42<br>Mkrn1 - makorin, ring finger protein, 1<br>Kdm5b - lysine (k)-specific demethylase 5b<br>Spp1 - secreted phosphoprotein 1<br>Mtf2 - metal response element binding transcription factor 2 |
| GO:1990830 | cellular response to leukemia inhibitory factor | 3.48E-04 | 6.45E-01 | 4.80 (900,32,41,7) | <br>Klf4 - kruppel-like factor 4 (gut)<br>Enah - enabled homolog (drosophila)<br>Zfp42 - zinc finger protein 42<br>Mkrn1 - makorin, ring finger protein, 1<br>Kdm5b - lysine (k)-specific demethylase 5b<br>Spp1 - secreted phosphoprotein 1<br>Mtf2 - metal response element binding transcription factor 2 |
| GO:1901360 | organic cyclic compound | 4.73E-04 | 7.00E-01 | 2.40 (900,137,41,15) | <br>Klf4 - kruppel-like factor 4 (gut) |

| | | | | | |
|---|---|---|---|---|---|
| | metabolic process | | | | Cdc5l - cell division cycle 5-like (s. pombe) |
| | | | | | Zfp57 - zinc finger protein 57 |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Hsd17b14 - hydroxysteroid (17-beta) dehydrogenase 14 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Nr0b1 - nuclear receptor subfamily 0, group b, member 1 |
| | | | | | Rpp25 - ribonuclease p/mrp 25 subunit |
| | | | | | Chchd10 - coiled-coil-helix-coiled-coil-helix domain containing 10 |
| | | | | | Tcea3 - transcription elongation factor a (sii), 3 |
| | | | | | H2afx - h2a histone family, member x |
| | | | | | Spp1 - secreted phosphoprotein 1 |
| | | | | | Upp1 - uridine phosphorylase 1 |
| | | | | | Dhx16 - deah (asp-glu-ala-his) box polypeptide 16 |
| | | | | | Morc1 - microrchidia 1 |
| GO:0070848 | response to growth factor | 5.73E-04 | 7.06E-01 | 5.27 (900,25,41,6) | |
| | | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Cdc5l - cell division cycle 5-like (s. pombe) |
| | | | | | Hspb1 - heat shock protein 1 |
| | | | | | Sox2 - sry-box containing gene 2 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Stmn2 - stathmin-like 2 |
| GO:0006304 | DNA modification | 9.73E-04 | 1.00E+00 | 7.98 (900,11,41,4) | |
| | | | | | Zfp57 - zinc finger protein 57 |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Morc1 - microrchidia 1 |
| **Function** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0003676 | nucleic acid binding | 9.11E-04 | 1.00E+00 | 2.03 (900,195,41,18) | |
| | | | | | Klf4 - kruppel-like factor 4 (gut) |
| | | | | | Cdc5l - cell division cycle 5-like (s. pombe) |
| | | | | | Zfp42 - zinc finger protein 42 |
| | | | | | Klf2 - kruppel-like factor 2 (lung) |
| | | | | | Zfp57 - zinc finger protein 57 |
| | | | | | Parp1 - poly (adp-ribose) polymerase family, member 1 |
| | | | | | Kdm5b - lysine (k)-specific demethylase 5b |
| | | | | | Tet1 - tet methylcytosine dioxygenase 1 |
| | | | | | Mtf2 - metal response element binding transcription factor 2 |
| | | | | | Dppa5a - developmental pluripotency associated 5a |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Nr0b1 - nuclear receptor subfamily 0, group b, member 1 |
| | | | | | Rpp25 - ribonuclease p/mrp 25 subunit |
| | | | | | Tcea3 - transcription elongation factor a (sii), 3 |
| | | | | | H2afx - h2a histone family, member x |
| | | | | | Sox2 - sry-box containing gene 2 |
| | | | | | Eif2s2 - eukaryotic translation initiation factor 2, subunit 2 (beta) |
| | | | | | Dhx16 - deah (asp-glu-ala-his) box polypeptide 16 |
| | | | | | Sod2 - superoxide dismutase 2, mitochondrial |
| **Component** | None | | | | |

Gene Ontology (GO) terms enriched for the top 50 gene loadings for MOFA factor 5, using 965 highly variable genes as a background:

| **Process** | | | | | |
|---|---|---|---|---|---|
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0072331 | signal transduction by p53 class mediator | 1.07E-06 | 7.88E-03 | 10.10 (900,13,48,7) | |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Foxo3 - forkhead box o3 |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Perp - perp, tp53 apoptosis effector |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| GO:0006974 | cellular response to DNA damage stimulus | 6.39E-06 | 2.36E-02 | 4.69 (900,44,48,11) | |
| | | | | | Btg2 - b cell translocation gene 2, anti-proliferative |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Foxo3 - forkhead box o3 |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | 1700007K13Rik - riken cdna 1700007k13 gene |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| | | | | | Ccng1 - cyclin g1 |
| | | | | | Marf1 - meiosis arrest female 1 |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| | | | | | Cbx3 - chromobox 3 |
| GO:0072332 | intrinsic apoptotic signaling pathway by p53 class mediator | 1.74E-05 | 4.30E-02 | 11.72 (900,8,48,5) | |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Perp - perp, tp53 apoptosis effector |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| GO:0006412 | translation | 2.09E-04 | 3.87E-01 | 7.81 (900,12,48,5) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| GO:0043043 | peptide biosynthetic process | 2.09E-04 | 3.10E-01 | 7.81 (900,12,48,5) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| GO:0030330 | DNA damage response, signal transduction by p53 class mediator | 2.22E-04 | 2.74E-01 | 10.71 (900,7,48,4) | |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Foxo3 - forkhead box o3 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| GO:0042771 | intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator | 2.22E-04 | 2.35E-01 | 10.71 (900,7,48,4) | |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Bbc3 - bcl2 binding component 3 |
| | | | | | Phlda3 - pleckstrin homology-like domain, family a, member 3 |
| GO:0042770 | signal transduction in response to DNA damage | 7.39E-04 | 6.84E-01 | 8.33 (900,9,48,4) | |
| | | | | | Cdkn1a - cyclin-dependent kinase inhibitor 1a (p21) |
| | | | | | Foxo3 - forkhead box o3 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Mdm2 - transformed mouse 3t3 cell double minute 2 |
| GO:0043604 | amide biosynthetic process | 9.81E-04 | 8.07E-01 | 5.86 (900,16,48,5) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| **Function** | | | | | |
| GO:0003735 | structural constituent of ribosome | 3.77E-05 | 5.19E-02 | 10.42 (900,9,48,5) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| **Component** | | | | | |
| **GO term** | **Description** | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
| GO:0022627 | cytosolic small ribosomal subunit | 3.44E-05 | 2.93E-02 | 15.00 (900,5,48,4) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| GO:0015935 | small ribosomal subunit | 3.44E-05 | 1.46E-02 | 15.00 (900,5,48,4) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Rps19 - ribosomal protein s19 |
| GO:0044391 | ribosomal subunit | 3.77E-05 | 1.07E-02 | 10.42 (900,9,48,5) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| GO:0005840 | ribosome | 1.27E-04 | 2.71E-02 | 8.52 (900,11,48,5) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |
| GO:0044445 | cytosolic part | 3.27E-04 | 5.56E-02 | 7.21 (900,13,48,5) | |
| | | | | | Rps4x - ribosomal protein s4, x-linked |
| | | | | | Rpl22l1 - ribosomal protein l22 like 1 |
| | | | | | Rps27l - ribosomal protein s27-like |
| | | | | | Rps23 - ribosomal protein s23 |
| | | | | | Rps19 - ribosomal protein s19 |

# Appendix I Gene names of ZGA signature list

The following list contains the gene names of the ZGA signature list used in chapter 5. The list was obtained after merging the gene names in Table S1 from Eckersley-Maslin et al. 2016, Table S8 from Hendrickson et al. 2017 and Table S1 from Y. Li et al. 2018.

0610005C13Rik, 0610009B14Rik, 0610031J06Rik, 0610040J01Rik, 1110005A03Rik, 1110006O24Rik, 1110017F19Rik, 1110018J18Rik, 1110021L09Rik, 1110032F04Rik, 1110038B12Rik, 1190002H23Rik, 1300014I06Rik, 1500010J02Rik, 1500012F01Rik, 1600002K03Rik, 1600010M07Rik, 1600012H06Rik, 1600015I10Rik, 1600020E01Rik, 1600025M17Rik, 1700001G17Rik, 1700007K13Rik, 1700009P17Rik, 1700010D01Rik, 1700012B15Rik, 1700013H16Rik, 1700016D06Rik, 1700019B21Rik, 1700019E08Rik, 1700019N12Rik, 1700024F13Rik, 1700025E21Rik, 1700029I01Rik, 1700034F02Rik, 1700034H15Rik, 1700048O20Rik, 1700060J05Rik, 1700069L16Rik, 1700080O16Rik, 1700084E18Rik, 1700086L19Rik, 1700086O06Rik, 1700093K21Rik, 1700096K18Rik, 1700112E06Rik, 1700123O20Rik, 1810019D21Rik, 1810026B05Rik, 1810032O08Rik, 1810035L17Rik, 1810044D09Rik, 1810062G17Rik, 2010001M09Rik, 2010204K13Rik, 2010317E24Rik, 2010320M18Rik, 2210016L21Rik, 2210404O07Rik, 2210414B05Rik, 2310003L22Rik, 2310011J03Rik, 2310040G24Rik, 2310045N01Rik, 2310047M10Rik, 2410002F23Rik, 2410004N09Rik, 2410016O06Rik, 2410075B13Rik, 2510002D24Rik, 2610005L07Rik, 2610019E17Rik, 2610027L16Rik, 2610028H24Rik, 2610206C17Rik, 2610306M01Rik, 2610318N02Rik, 2700023E23Rik, 2700038G22Rik, 2700078E11Rik, 2810004N23Rik, 2810006K23Rik, 2810008D09Rik, 2810021B07Rik, 2810029C07Rik, 2810055F11Rik, 2810405K02Rik, 2810417H13Rik, 2810429I04Rik, 2810459M11Rik, 2900002K06Rik, 2900079G21Rik, 3010003L10Rik, 3110056K07Rik, 3830408C21Rik, 4632427E13Rik, 4732471D19Rik, 4831440E17Rik, 4921517L17Rik, 4921531C22Rik, 4930413G21Rik, 4930427A07Rik, 4930447C04Rik, 4930455C21Rik, 4930465K10Rik, 4930479M11Rik, 4930483J18Rik, 4930515G01Rik, 4930528A17Rik, 4930547E08Rik, 4930558C23Rik, 4930578G10Rik, 4930578I07Rik, 4930579G24Rik, 4930583H14Rik, 4932411N23Rik, 4933404O12Rik, 4933406J08Rik, 4933408N05Rik, 4933411G11Rik, 4933430I17Rik, 4933430M04Rik, 4933440M02Rik, 5330411J11Rik, 5330426P16Rik, 5430402O13Rik, 5430416N02Rik, 5730408K05Rik, 5730419I09Rik, 5730455P16Rik, 5730508B09Rik, 5730590G19Rik, 5830403L16Rik, 5830433M19Rik, 6030408B16Rik, 6230400D17Rik, 6230427J02Rik, 6430573F11Rik, 9130008F23Rik, 9330020H09Rik, 9330133O14Rik, 9330159M07Rik, 9430008C03Rik, 9430015G10Rik, 9430016H08Rik, 9430021M05Rik, 9430060I03Rik, 9530077C05Rik, A130040M12Rik, A2m, A330049M08Rik, A430035B10Rik, A430084P05Rik, A430089I19Rik, A530032D15Rik, A530040E14Rik, A630001G21Rik, A630066F11Rik, A630072M18Rik, A730018C14Rik, A730037C10Rik, A730085A09Rik, AA467197, Aacs, Abcb10, Abcb5, Abcc8, Abhd14b, Abhd3, Ablim1, Abo, Abpg, AC079644.1, AC079644.2, AC079644.3, AC091683.1, AC091683.2, AC121866.1, AC121888.1, AC122464.1, AC122464.2, AC124724.1, AC127274.2, AC127374.1, AC130840.1, AC132362.1,

AC134841.1, AC140240.1, AC140409.1, AC153539.1, AC153998.1, AC156643.1, AC158600.1, AC158600.2, AC158600.3, AC158600.4, AC158600.6, AC164627.1, AC165327.2, AC211878.1, AC231112.1, AC231112.2, AC238940.1, AC238940.3, AC239617.1, AC239678.2, AC240744.3, Acaa1a, Acad8, Acap3, Acn9, Acot1, Acrbp, Acrv1, Acss3, Actr2, Actr5, Adad1, Adam2, Adamts1, Adamts5, Adck4, Adcy2, Adh7, Adi1, Adm, Adnp, Adora2b, Adrb3, Adrbk1, AF067061, AF067063, Aff1, Aga, Agmat, Agpat9, Agrp, Agtrap, Ahnak, Ahnak2, AI427809, AI462493, AI464131, AI838599, Ajap1, Ak4, Akr1b7, Aldh18a1, Aldh1l1, Alg5, Alkbh4, Aloxe3, Alppl2, Amd2, Amhr2, Amica1, Ankdd1b, Ankrd17, Ankrd22, Ankrd35, Ankrd50, Ankrd54, Ankrd9, Antxr1, Ap2a2, Ap4m1, Apoa2, Apoc2, Apod, Apol7b, Apom, Aqp11, Aqp3, Aqr, Arap2, Arf2, Arfgap2, Arfip2, Arfrp1, Arg1, Arg2, Arhgap29, Arhgap36, Arhgap8, Arhgdib, Arhgdig, Arhgef26, Arid3c, Arid5a, Arid5b, Arih2, Arl13b, Arl14, Arl15, Arl16, Arl5c, Arnt, Arrdc3, Arrdc4, Arx, Ascl1, Ascl2, Asl, Asns, Ate1, Atf2, Atf3, Atf5, Atg4d, Atg9a, Atp5s, Atp6v1e1, Atp6v1g3, Atp8b1, Atxn7l3, AU019990, AU022252, AU041133, Auts2, Avpi1, AW822073, Axin2, AY761184, B020004C17Rik, B020004J07Rik, B020031M17Rik, B3gnt4, B3gntl1, B4galnt2, B4galt1, B4galt6, B930059L03Rik, Baat1, Bag3, Bambi, Bambips1, BB287469, "BB287469, Gm4027", BC002230, BC003965, BC016495, BC017647, BC025920, BC027231, BC028528, BC033916, BC037704, BC046404, BC048355, BC048507, BC049715, BC049762, BC057022, BC061212, BC080695, BC147527, Bcdin3d, Bcl2l14, Bcor, Bcorl1, Bcs1l, Bdh1, Bdnf, Bend3, Bend6, Best2, Bex1, Bex2, Bex6, Bhlhb9, Bicc1, Bid, Blnk, Blvrb, Bmp2, Bmyc, Bnc2, Bnip3, Bola2, Boll, Bop1, Brd2, Bre, Bri3bp, Bst1, Btbd19, Btg1, Btg2, Bud13, C030034I22Rik, C130026I21Rik, C130060C02Rik, C130074G19Rik, C1d, C1ql1, C1qtnf4, C230096C10Rik, C2cd2l, C2cd4b, C630043F03Rik, C86695, Cab39, Cacna1h, Cacna1s, Calcoco2, Cald1, Calml4, Cand1, Cap1, Capn5, Capsl, Car6, Caskin2, Catsperg1, Cbx7, Ccdc106, Ccdc110, Ccdc116, Ccdc125, Ccdc126, Ccdc134, Ccdc137, Ccdc160, Ccdc50, Ccdc60, Ccdc64b, Ccdc66, Ccdc83, Ccdc85b, Cchcr1, Ccl3, Ccng2, Ccnjl, Ccno, Ccrn4l, Ccs, Cct8l1, Cd200r2, Cd24a, Cd52, Cd63, Cd97, Cda, Cdan1, Cdc42ep3, Cdc42ep4, Cdk4, Cdk5r1, Cdk5rap3, Cdkn2aip, Cdkn3, Cdrt4, Cdsn, Cdyl2, Ceacam1, Ceacam19, Cebpa, Celf4, Cep57l1, Cep97, Cetn1, Chchd4, Chchd5, Chek2, Chga, Chit1, Chkb, Chrac1, Chrm3, Chrna10, Chrna5, Chrna9, Chst7, Chtf18, Chtf8, Churc1, Cirbp, Cited1, Cited4, Clcn5, Cldn18, Cldn3, Cldn5, Cldn6, Clec10a, Clic1, Clip2, Clk1, Clk3, Cln8, Clp1, Clu, Cml1, Cml2, Cmtm2a, Cndp2, Cnn2, Cnnm2, Cnnm3, Cnpy1, Cnpy3, Cntnap3, Cobl, Coch, Col12a1, Col13a1, Col28a1, Col4a1, Col4a2, Colq, Commd7, Coq4, Cotl1, Cox19, Cox8c, Cpb2, Cpe, Cpeb2, Cpt1a, Crabp1, Crebl2, Creld1, Crip1, Crtap, Crx, Crxos1, Csmd1, Csrnp1, Csrnp3, Cst6, Cst7, Cst9, Cstb, CT025616.1, CT030687.1, CT485612.1, CT485612.2, CT954323.2, Ctf2, Ctla2b, Ctr9, Ctrb1, Ctsa, Ctsc, Ctsl, Ctss, Ctsz, Ctu1, Cuedc2, Cwc22, Cwc25, Cxadr, Cxcl10, Cxcl11, Cxcl16, Cyba, Cyp2b23, Cyp2c44, Cyp2c67, Cyp2s1, "Cypt1, Cypt8", Cypt12, Cypt2ps, Cypt3, "Cypt4, Cypt9", Cypt7, D17Ertd648e, D1Pas1, D4Wsu53e, D5Ertd605e, D7Ertd143e, Dact3, Dapl1, Dars2, Dazl, Dbc1, Dbndd2, Dbnl, Dbp, Dbr1, Dbx1, Dcbld1, Dcc, Dcdc5, Dclre1c, Dcp2, Ddhd1, Ddit4, Ddit4l, Ddr2, Ddx26b, Ddx31, Ddx39, Ddx43, Decr2, Dedd2, Def8, Defb13, Defb23, Defb25, Dennd4c, Depdc1b, Depdc5, Depdc7, Dexi, Dgkk, Dhcr24, Dhcr7, Dhdds, Dhh, Dhrs7b, Dhtkd1, Dido1, Dkkl1, Dlgap2, Dlgap3, Dlx2, Dmrt1, Dnahc7b, Dnajb14, Dnajb3, Dnajb9, Dnajc12, Dnajc28, Dnajc6, Dnlz, Dnpep, Doc2a, Dock9, Dohh, Dolpp1, Dpagt1, Dpep1, Dpf3, Dph2, Dpp7, Dppa2, Dppa3, Dppa5a, Dpys, Dpysl3, Dpysl5, Drd3, Drr1, Dsg1b, Dst, Dub1, Dub1a, Dub2a, Dub3, Duox2, Duoxa2, Dusp1, Dusp28, Dusp4, Dux, Dyrk3, Dyrk4, Dzip1, E130309D02Rik, E2f7, Ears2, Ebf3, Ecel1, Echdc2, Echdc3, Ecm2, Ecsit, Efcab4b, Egflam, Egr1, Eid2, Eif2s3y, Eif4ebp3, Eif5a2, Elovl4, Elovl6, Emp3, Endod1, Enpp3, Entpd7, Epha10, Ephb1, Epm2a, Epm2aip1, Ercc4, Ero1l, Errfi1, Esp24, Esrrb, Esrrg, Esyt3, Etfdh, Etohd2, Etv3, Etv5, Exoc7, Exoc8, Exog, Exosc6, F11r, Fabp9, Fadd, Fah, Fahd1, Fam102b, Fam105a, Fam109a, Fam124a, Fam134a, Fam13c, Fam150a, Fam151a, Fam158a, Fam171b, Fam173b, Fam176a, Fam181b, Fam190a, Fam195b, Fam203a, Fam20b, Fam33a, Fam43b, Fam53c, Fam57b, Fam73a, Fam76a, Fam81a, Fam84a, Fam84b, Fam89b, Fan1, Fancf, Fat3, Fblim1, Fbll1, Fbxl15, Fbxo15, Fbxo31, Fbxo34, Fbxo6, Fbxw9, Fcgbp, Fcgr2b, Fcgrt, Fdxr, Fgf1, Fgf4, Filip1l, Fkbp11, Fkbp1b, Flrt2, Flt4, Flywch2, Folr1, Folr2, Fos, Foxi3, Foxo6, Foxp1, Foxred1,

Frat2, Frrs1, Fscn1, Fstl4, Fthl17, Fundc1, Fv1, Fxyd6, Fzd4, Fzd5, Fzd7, Fzd9, G2e3, Gabarap, Gabpa, Gabra1, Gadd45g, Gal, Gal3st2, Galns, Galnt3, Galt, Gan, Gata1, Gba2, Gcdh, Gcnt1, Gcnt2, Gdap10, Gdpd4, Gfra3, Ggn, Gja1, Gjb3, Gla, Glipr2, Glod5, Glrx, Glrx2, Glrx3, Glt25d1, Glud1, Glul, Gm10188, Gm10264, Gm10354, Gm10394, Gm10424, Gm10505, Gm10553, Gm10639, Gm10668, Gm10696, Gm10718, Gm10800, Gm10801, Gm11052, Gm11232, Gm11236, Gm11237, Gm11238, Gm11239, Gm11487, Gm11517, Gm11544, Gm11564, Gm11602, Gm11640, Gm11756, Gm11757, Gm11974, Gm12088, Gm12114, Gm12531, Gm12702, Gm12714, Gm12724, Gm12730, Gm12789, Gm12790, Gm12794, Gm12800, Gm12823, Gm12824, Gm12953, Gm13040, Gm13043, Gm13057, Gm13078, Gm13083, Gm13101, Gm13109, Gm13119, Gm13128, Gm13139, Gm13335, Gm13498, Gm13693, Gm13694, Gm13695, Gm13696, Gm13698, Gm13718, Gm13871, Gm13962, Gm13964, Gm14322, Gm14325, Gm14326, Gm14393, Gm14403, Gm14634, Gm14742, Gm14798, Gm14929, Gm15023, Gm1527, Gm15421, Gm15455, Gm15787, Gm16008, Gm16023, Gm16028, Gm16062, Gm16119, Gm16211, Gm16239, Gm16243, Gm16381, Gm16429, Gm16513, Gm16517, Gm17019, Gm17026, Gm17611, Gm1995, Gm2016, Gm20199, Gm2022, Gm2027, Gm2042, Gm20431, Gm20440, Gm2046, Gm20580, Gm20625, Gm20631, Gm20634, Gm2075, Gm281, Gm2a, Gm3139, Gm3258, Gm4027, Gm428, Gm4301, Gm4302, Gm4303, Gm4305, Gm4307, Gm4312, Gm4340, Gm44, Gm4532, Gm4778, Gm4782, Gm4827, Gm4858, Gm4971, Gm498, Gm4981, Gm4984, Gm5077, Gm5127, Gm5148, Gm5286, Gm5577, Gm5590, Gm5612, Gm5635, Gm5647, Gm5662, Gm5698, Gm5699, Gm5773, Gm581, Gm6086, Gm6189, Gm6351, Gm6432, Gm6468, Gm6502, Gm6507, Gm6509, Gm6568, Gm6654, Gm6763, Gm6880, Gm6890, Gm6902, Gm7102, Gm749, Gm7647, Gm7682, Gm773, Gm7942, Gm7982, Gm8038, Gm8094, Gm8104, Gm8300, Gm8766, Gm8994, Gm9, Gm9116, Gm9125, Gm973, Gm9895, Gm9958, Gmeb2, Gmpr2, Gna15, Gnat1, Gnaz, Gnb4, Gnl1, Gnl3l, Golga1, Gorasp1, Gpa33, Gpatch3, Gpbp1l1, Gpc5, Gpkow, Gpr126, Gpr137c, Gpr161, Gpr182, Gpr19, Gpr50, Gpr63, Gpr75, Gpr83, Gprc5c, Gpx7, Grb7, Gria1, Grifin, Grik3, Grk4, Grk6, Grp, Grwd1, Gsta4, Gstm1, Gstm6, Gsto1, Gt(ROSA)26Sor, Gtf2b, Gtf2i, Gtf3c4, Gtf3c5, Gtpbp3, Gtsf1l, Guca1a, Gulo, Gusb, H1f0, H2Bl, H2D1, H2DMa, H2M10.4, "H2Q7, H2Q9", H2Q8, "H2T22, H2T9", H2T9, H2afx, H47, Hadha, Hars2, Hbaa2, Hcrt, Hdc, Hddc3, Hdhd3, Herpud1, Hes1, Hesx1, Hexa, Hexb, Hexdc, Hexim1, Hinfp, Hip1, Hipk1, Hist1h1a, Hist1h1c, Hist1h2aa, Hist1h2ab, Hist1h2ac, Hist1h2ad, Hist1h2ag, Hist1h2ah, Hist1h2bh, Hist1h2bj, Hist1h2bk, Hist1h2bl, Hist1h2bp, Hist1h3c, Hist1h3d, Hist1h3e, Hist1h3g, Hist1h3h, Hist1h4b, Hist1h4f, Hist1h4i, Hist1h4j, Hist1h4n, Hist2h3c2, Hist3h2ba, Hlx, Hmgb4, Hmgcl, Hmgn3, Hmx1, Hoxa1, Hoxa9, Hoxb2, Hoxd10, Hoxd11, Hoxd13, Hpdl, Hrh2, Hs6st2, Hsd17b10, Hsd17b14, Hspa1a, Hspa1b, Hspa1l, Hspa2, Hspa8, Hspb3, Hspb8, Hspbp1, Htra1, I0C0044D17Rik, Iah1, Iars, Id1, Id3, Id4, Idh2, Idh3b, Ier2, Ier3, Ier3ip1, Ier5, Iffo1, Ifi35, Ifitm1, Ifitm3, Ifltd1, Ifng, Igf2bp1, Igfbp2, Igfbp3, Ighe, Igtp, Ikbip, Ikzf5, Il12rb2, Il18rap, Il2rg, Il6ra, Il6st, Impa1, Impdh1, Ing4, Inpp4b, Inpp5j, Insl6, Insrr, Ints7, Ipo4, Iqcf1, Iqch, Iqub, Irak2, Irak3, Irf2bp1, Irf3, Irf7, Irf9, Irgq, Irx1, Irx2, Irx4, Isca2, Isg15, Isg20, Isg20l2, Itfg2, Itgae, Itpkc, Jam2, Jmjd4, Jmjd8, Jmy, Jub, Jun, Junb, Kank4, Kat2b, Katnb1, Kbtbd10, Kbtbd2, Kcna1, Kcna3, Kcne1, Kcnj13, Kcnk6, Kcnn2, Kdelc2, Kdm4c, Kdm5a, Kif14, Kif18b, Kifc2, Kirrel, Klb, Klf3, Klf5, Klf9, Klhl13, Klhl7, Klrg2, Kpna1, Kpna4, Kri1, Krt18, Krt28, Krtap413, L1td1, L2hgdh, Laptm5, Lass5, Lass6, Lat2, Lck, Lctl, Ldhc, Lefty2, Lemd3, Leng8, Lgals1, Lgals12, Lgals2, Lgals4, Lgals9, Lhfpl1, Lias, Limch1, Lime1, Limk1, Lin7a, Lin7b, Lmbr1l, Lmo4, Lmx1a, Lnp, LOC100503496, Lonp2, Lonrf3, Loxl1, Lpar6, Lpcat3, Lpcat4, Lphn2, Lphn3, Lrig1, Lrp12, Lrrc2, Lrrc38, Lrrc3b, Lrrn4, Lst1, Lta, Lxn, Ly6g5b, Lyrm1, Lyrm2, Lyrm5, Macrod1, Mad2l1bp, Mafb, Mafk, Magel2, Mak16, Malt1, Man1b1, Man2a2, Man2b1, Maneal, Map2k3, Map3k3, Mapkapk2, March1, March3, Marcksl1, Mars2, Marveld1, Maz, Mb, Mbd4, Mbd5, Mbd6, Mbnl3, Mccc2, Mcm6, Mcm9, Mctp2, Mcts2, Mdn1, Mdp1, Mecom, Med26, Med29, Mef2a, Mef2d, Meg3, Meox2, Mep1b, Mesdc2, Mest, Mettl13, Mettl2, Mfap4, Mfap5, Mfsd12, Mfsd5, Mfsd7b, Mfsd7c, Mga, Mgat2, Mgll, Micall1, Micu1, Mir17hg, Mlh1, Mlh3, Mlxip, Mmp19, Mmrn2, Mn1, Mob3a, Mob3b, Mob3c, Morc1, Mpdu1, Mpp1, Mppe1, Mpv17, Mreg, Mrgprb1, Mrgprx2, Mrm1, Mrpl17, Mrpl33, Mrps34, Msra, Msx1, Mt1, Mt2, Mt3, Mta2, Mtap6, Mtap7d3, Mtch2, Mterfd3, Mthfr, Mtrf1l, Mtss1, Mttp, Muc13,

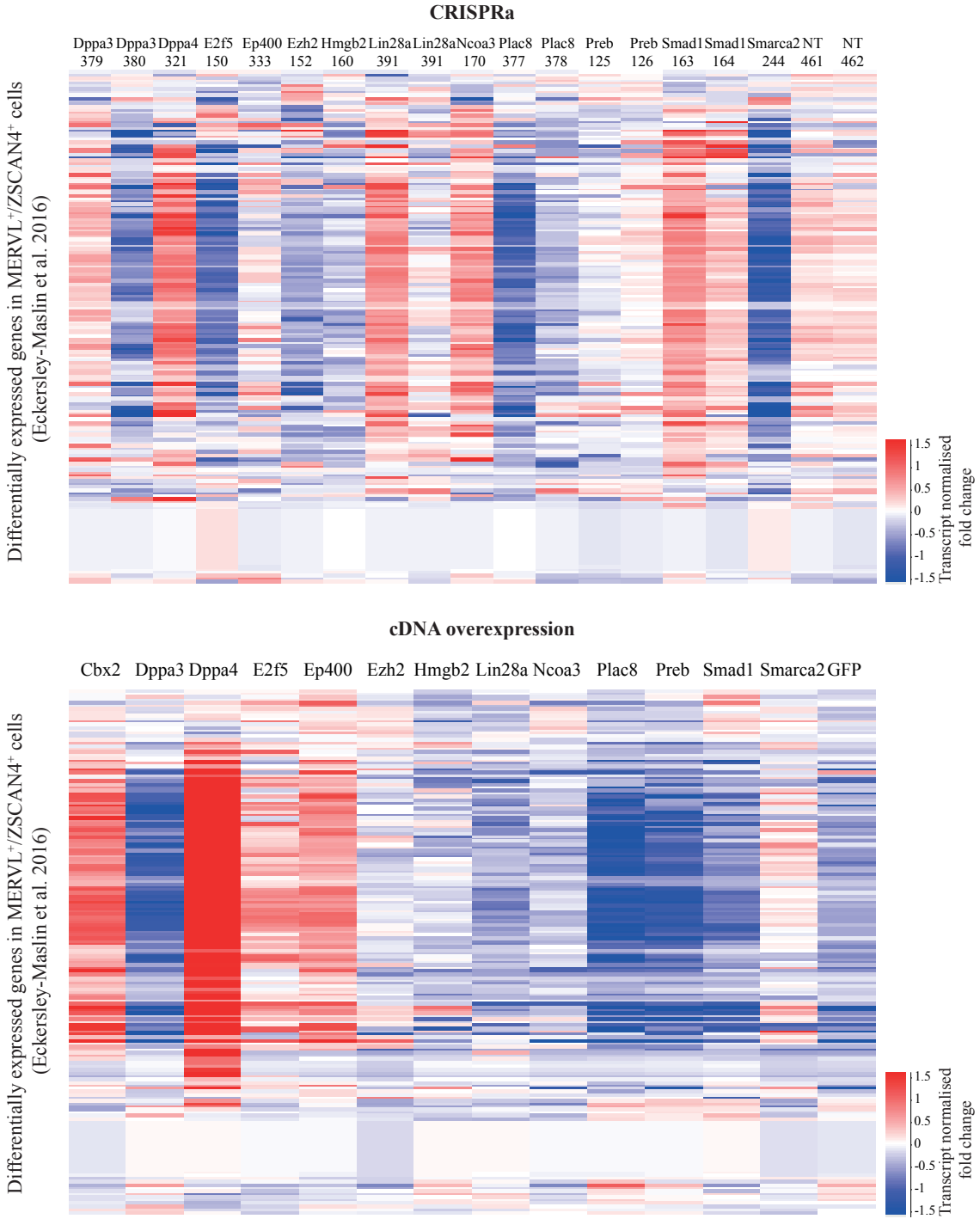
Mum1l1, Mvd, Mvk, Myadml2, Myc, Mycs, Myg1, Myl3, Mylpf, Myo1e, Myof, N4bp3, N6amt2, Naalad2, Nanog, Nanos3, Nanp, Narfl, Ncbp1, Ncf1, Ndrg1, Ndufaf1, Ndufaf3, Ndufb7, Ndufc1, Nedd4, Nedd4l, Nefl, Nek10, Neto2, Neu4, Neurl2, Neurog2, Nfam1, Nfat5, Nfatc3, Ngfrap1, Nid1, Nid2, Ninj2, Nipsnap1, Nkapl, Nkx25, Nlrp4f, Nme4, Nmnat2, Nop16, Nop2, Npc1l1, Npl, Nr2c2, Nrarp, Nrip1, Nrip3, Nrp, Nrxn3, Nsmaf, Nt5dc2, Nudt16l1, Nudt22, Nup62cl, Nupr1, Nxf2, Oaz3, Obox1, Obox2, Obox3, Obox6, Olfm3, Olfr118, Olfr119, Olfr120, Olfr1277, Olfr161, Olfr18, Olfr214, Olfr293, Olfr328, Olfr376, Olfr450, Olfr697, Olfr699, Olfr787, Olfr788, Olfr815, Olfr847, Olfr881, Oog4, Oraov1, Ormdl3, Os9, Osbpl6, Osgep, Osgin2, Osm, Osr2, Otud6a, Otx1, Ovol1, Ovol2, Oxsr1, P2ry1, P4ha2, Paip2b, Pank3, Papolb, Papolg, Parp10, Pax9, Pcdh10, Pcdh17, Pcdh8, Pcdhgb7, Pcf11, Pclo, Pcyt2, Pdcd1, Pde3b, Pde6a, Pde9a, Pdgfrl, Pdk3, Pdk4, Pdlim3, Pdp2, Pdxk, Pemt, Per1, Pgap1, Pgk2, Pgm3, Phc2, Phf11, Phf16, Phf17, Phf21b, Phldb2, Pi4k2b, Pias3, Pif1, Pigl, Pigo, Pin1, Pisdps1, Pitpnb, Pitpnc1, Piwil2, Pla2g16, Pla2g1b, Plcb1, Plcd4, Plcl1, Pld4, Plekhf1, Plekhm1, Plekho2, Plk2, Plk3, Plod3, Plp2, Plxdc2, Pmaip1, Pmm2, Pmpca, Pnp, Pnp2, Polg2, Polh, Polq, Polr2a, Pomc, Popdc3, Porcn, Pou2f3, Pou3f1, Pou6f2, Ppat, Ppcs, Ppig, Ppil1, Ppp1r12c, Ppp1r15a, Ppp1r2ps7, Ppp1r27, Ppp1r8, Ppp2cb, Ppp2r3a, Pramef6, Pramef8, Pramel1, Pramel3, Pramel4, Pramel5, Pramel6, Pramel7, Prdm15, Prdm9, Prdx4, Prex2, Prkch, Prkcz, Prkg1, Prmt10, Prmt6, Prnp, Prps1, Prr14, Prr15l, Prr19, Prrg2, Prss23, Prss8, Prtg, Prtn3, Psat1, Psmb10, Psmd13, Psme1, Psmg2, Ptch1, Ptpmt1, Ptpn18, Ptpn4, Ptprg, Ptrf, Ptrh1, Pura, Purb, Purg, Pusl1, Pygb, Pygm, Qpctl, Qrsl1, Qtrt1, Rab11fip4, Rab19, Rab20, Rab24, Rab30, Rab34, Rab39, Rab40c, Rab42ps, Rab43, Rab4b, Rab7l1, Rabep2, Rabepk, Rad9, Rad9b, Radil, Ramp1, Ranbp6, Rapgef2, Rarg, Rasd1, Rasgef1c, Rasl11a, Rasl29, Rbm10, Rbm15, Rbm41, Rbms3, Rbmx, Rbp7, Rc3h2, Rccd1, Rcn3, Rdh1, Rdh16, Rdh9, Rdm1, Recql4, Reep6, Renbp, Rfc5, Rfpl4b, Rfx4, Rg9mtd2, Rgl2, Rgn, Rgs11, Rgs16, Rgs2, Rhbdl2, Rhox5, Ric3, Rilpl2, Rimbp2, Rimklb, Ripk4, Rln1, Rnase4, Rnaseh2a, Rnf103, Rnf113a1, Rnf121, Rnf128, Rnf19b, Rnmtl1, Rnpc3, Robo1, Rpgrip1, Rpia, Rpl10l, Rpl39l, Rpp25, Rpp40, Rprd1b, Rps26, Rps27, Rpusd1, Rpusd2, Rrp9, Rrs1, Rsph1, Rtn2, Rundc1, Rusc1, Rxra, Rxrg, Sall4, Sat2, Satb2, Scamp1, Sccpdh, Scd1, Scd2, Scg5, Sco1, Sco2, Sdr39u1, Sdr9c7, Sec1, Sec61a2, Sel1l2, Sephs2, Sept1, Sept11, Sept2, Sept6, Sept9, Sepx1, Serpinb1c, Serpinf1, Sertad1, Sesn1, Sfta2, Sfxn2, Sfxn4, Sgk3, Sh2d3c, Sh3bp4, Sh3kbp1, Shb, Shroom1, Siah2, Sik1, Six1, Six4, Skil, Slc10a3, Slc14a2, Slc16a3, Slc16a6, Slc1a4, Slc1a5, Slc20a1, Slc22a20, Slc22a28, Slc25a14, Slc25a43, Slc26a1, Slc27a2, Slc27a5, Slc29a2, Slc2a1, Slc2a8, Slc30a2, Slc34a2, Slc35e2, Slc35e3, Slc38a2, Slc38a4, Slc39a1, Slc39a14, Slc39a4, Slc4a4, Slc4a5, Slc52a2, Slc5a4b, Slc5a6, Slc6a6, Slc6a8, Slc7a3, Slc7a9, Slc9a9, Slfn10ps, Smad7, Smg5, Smpd3, Smtn, Smyd5, Snai1, Snai2, Snai3, Snhg11, Snhg12, Snhg3, Snhg4, Snhg7, Snhg8, Snrnp35, Snrpc, Snta1, Snw1, Snx8, Socs2, Socs3, Sord, Sowaha, Sowahc, Sox10, Sox11, Sox2, Sox3, Sox30, Sox8, Sp110, Sp140, Sp4, Sp6, Spag9, Spdya, Specc1, Speer4c, Speer4d, Speer4e, Speer7ps1, Speer8ps1, Spesp1, Spic, Spin2, Spns1, Spock3, Spp1, Sprr2d, Sprr2e, Spry2, Spryd4, Spty2d1, Spz1, Srd5a1, Srgap3, Srl, Stac2, Stag3, Stard6, Steap1, Stim1, Stk16, Stk17b, Stk19, Stk38, Stmn3, Ston2, Stox1, Sult2b1, Sult5a1, Sun1, Surf2, Sva, Sycp1ps1, Syde2, Syne2, Syngr1, Synm, Synpo2, Sytl2, Tacc3, Tacr3, Taf1d, Tagln, Tagln2, Tapbp, Tatdn3, Tbc1d12, Tbl1x, Tbl2, Tbl3, Tbrg3, Tbrg4, Tbx1, Tbx20, Tbx3, Tbxa2r, Tceal8, Tchh, Tcirg1, Tcstv1, Tcstv3, Tdh, Tdpoz1, Tdpoz2, Tdpoz3, Tdpoz4, Tdpoz5, Tef, Tekt2, Terc, Tex101, Tex13, Tex19.1, Tfap2a, Tfcp2, Tfcp2l1, Tfpi2, Tfrc, Thbs1, Thnsl2, Thsd7b, Thy1, Ticam2, Tie1, Timd2, Timm22, Tktl1, Tle3, Tlk2, Tm4sf1, Tmco2, Tmed6, Tmeff1, Tmem101, Tmem106a, Tmem119, Tmem126b, Tmem129, Tmem146, Tmem167, Tmem167b, Tmem177, Tmem189, Tmem191c, Tmem20, Tmem208, Tmem215, Tmem229b, Tmem35, Tmem37, Tmem39a, Tmem56, Tmem79, Tmem92, Tmppe, Tmsb15b1, Tmsb15b2, Tmsb15l, Tmx1, Tmx2, Tnfaip6, Tnfaip8l2, Tnfrsf13b, Tnfrsf17, Tnip2, Tnnc2, Tob1, Toe1, Tomm5, Tor1b, Tor3a, Tpo, Traf2, Trak2, Trap1a, Trh, Trib3, Trim25, Trim32, Trim43a, Trim43b, Trim43c, Trim52, Trim8, Triml2, Trmt1, Trmt61a, Trmt61b, Trp53bp2, Trub2, Tsc22d3, Tsen2, Tsen54, Tsga8, Tspan1, Tspan4, Tspan6, Tssk6, Tsx, Ttc30b, Ttc39d, Ttll12, Tuba1a, Tubb2b, Txnip, Uap1, Uap1l1, Ubap1, Ubc, Ube2o, Ube2t, Ube2w, Ubtd2, Ubtfl1, Ubxn2a, Ugt1a1, Uhrf1bp1, Uhrf2, Uimc1, Unc119b, Upk1a, Upk3a, Uqcrc1, Ush1g, Uspps, Usp17l5, Usp25, Usp26, Usp50, Usp9y, Utf1,

Utp14b, Utp23, Uty, Vcam1, Vcan, Vegfa, Vgll1, Vhl, Vim, Vmn1r227, Vmn1r53, Vmn1r88, Vmn2r1, Vmn2r94, Vps33a, Vps39, Wbscr27, Wdr16, Wdr18, Wdr27, Wdr4, Wdr45, Wdr62, Wdr77, Wdtc1, Whsc2, Wipf2, Wnt5a, Wrap53, Wt1, Wwtr1, Xab2, Xaf1, Xkr9, Xlr, Xpnpep1, Xpot, Xylb, Yars, Ydjc, Yeats2, Yif1a, Yod1, Zbed6, Zbtb17, Zbtb5, Zbtb7a, Zbtb8a, Zc3h10, Zc3h12a, Zc3h7a, Zc3hc1, Zcchc11, Zcchc12, Zcchc13, Zcchc17, Zcchc24, Zfand2a, Zfp1, Zfp119b, Zfp142, Zfp184, Zfp187, Zfp217, Zfp239, Zfp280c, Zfp292, Zfp296, Zfp30, Zfp317, Zfp335, Zfp352, Zfp353, Zfp367, Zfp382, Zfp386, Zfp42, Zfp446, Zfp51, Zfp516, Zfp53, Zfp54, Zfp560, Zfp566, Zfp57, Zfp572, Zfp574, Zfp593, Zfp599, Zfp608, Zfp622, Zfp623, Zfp637, Zfp647, Zfp689, Zfp704, Zfp706, Zfp707, Zfp719, Zfp771, Zfp775, Zfp800, Zfp808, Zfp809, Zfp810, Zfp825, Zfp867, Zfp871, Zfp874b, Zfp882, Zfp941, Zfp948, Zfpm2, Zfx, Zfy1, Zfy2, Zfyve26, Zgpat, Zic3, Zkscan1, Zmiz1, Zmym2, Znhit2, Znhit3, Zrsr1, Zscan4a, Zscan4b, Zscan4c, Zscan4d, Zscan4e, Zscan4f, Zswim2, Zswim3, Zswim5, Zswim6, Zyg11a, Zyg11b

# Appendix J Validation of additional screen candidates

During initial analysis of the CRISPRa scRNA-seq screen dataset, before developing the MOFA model in collaboration with Danila Bredikin and Oliver Stegle (EMBL, Heidelberg, see Table of Acknowledgement of Assistance), I identified potential hits by differential gene expression analysis of effective sgRNAs. Briefly, using an FDR thereshold of <10%, I called differentially expressed genes (DEG) and, subsequently, calculated the fraction of upregulated ZGA genes considering those previously reported in published literature (Eckersley-Maslin et al. 2016; Hendrickson et al. 2017; Y. Li et al. 2018) (Appendix I). As shown in the following table, 29 sgRNAs were ranked for differentially upregulating at least one ZGA gene, and some of them, in addition to *Dppa2*, *Smarca5*, and *Patz1*, were validated by CRISPRa and cDNA overexpression, as described in Figure 5.17. *Dppa4* was included as an additional positive control (Eckersley-Maslin et al. 2019, De Iaco et al. 2019) and *Plac8* and *Preb* were included as negative controls for inducing target gene activation but not upregulating ZGA genes (Appendix C), similar to the *Carhsp1* control used in chapter 5. The results of these validations are presented here. However, the low number of differentially expressed genes captured suggested this approach was less than ideal to call screen hits and, therefore, most of these candidates were not analysed further. Instead, MOFA was developed to overcome the sensitivity limitations of scRNA-seq data and identify screen hits that induced a ZGA signature, rather than individual ZGA genes.

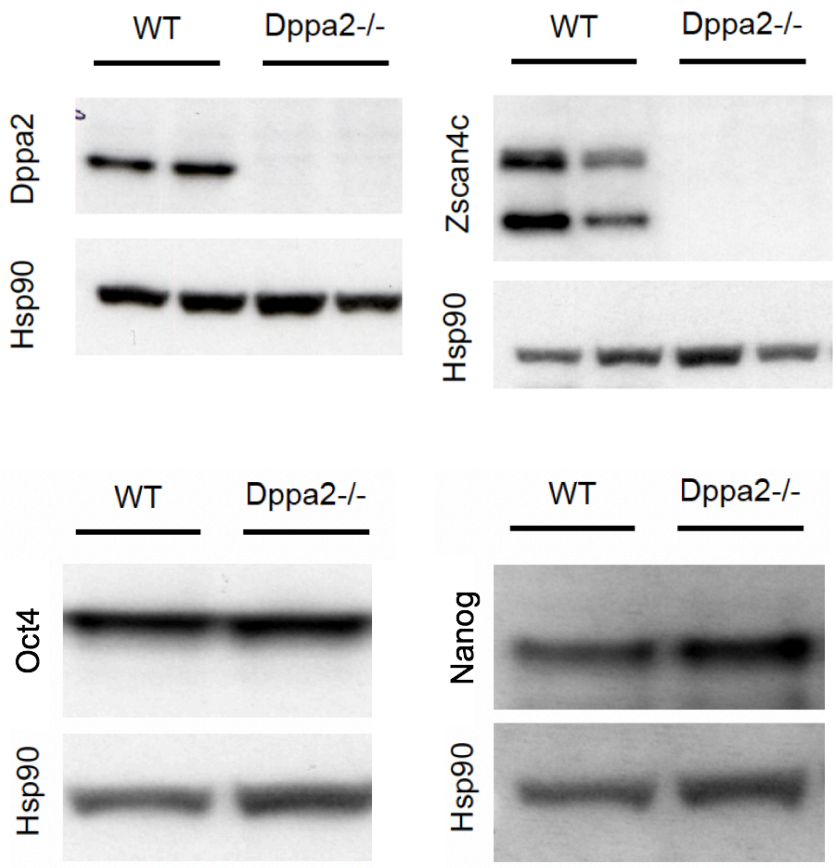
| sgRNA ID | Target gene | N of DEG (FDR 10%) | N of upregulated genes (FDR 10%) | N of ZGA genes (FDR 10%) | N of upregulated ZGA genes (FDR 10%) | Fraction of upregulated ZGA genes |
|---|---|---|---|---|---|---|
| 323 | *Dppa2* | 52 | 28 | 12 | 11 | 0.392857143 |
| 353 | *Smarca5* | 111 | 54 | 19 | 15 | 0.277777778 |
| 160 | *Hmgb2* | 22 | 15 | 4 | 3 | 0.2 |
| 152 | *Ezh2* | 72 | 60 | 11 | 11 | 0.183333333 |
| 379 | *Dppa3* | 296 | 147 | 26 | 21 | 0.142857143 |
| 270 | *Smarce1* | 20 | 14 | 2 | 2 | 0.142857143 |
| 150 | *E2f5* | 282 | 133 | 35 | 16 | 0.120300752 |
| 103 | *Patz1* | 246 | 123 | 20 | 14 | 0.113821138 |
| 153 | *Gata3* | 87 | 73 | 8 | 8 | 0.109589041 |
| 455 | *Zscan29* | 62 | 46 | 8 | 5 | 0.108695652 |
| 163 | *Smad1* | 61 | 47 | 7 | 5 | 0.106382979 |
| 74 | *Sirt1* | 23 | 20 | 2 | 2 | 0.1 |
| 211 | *Uhrf1* | 11 | 11 | 1 | 1 | 0.090909091 |
| 396 | *Chd3* | 32 | 25 | 3 | 2 | 0.08 |
| 242 | *Pou5f1* | 734 | 461 | 60 | 36 | 0.078091106 |
| 333 | *Ep400* | 51 | 41 | 7 | 3 | 0.073170732 |
| 170 | *Ncoa3* | 39 | 28 | 5 | 2 | 0.071428571 |
| 391 | *Lin28a* | 92 | 79 | 6 | 5 | 0.063291139 |
| 69 | *Max* | 18 | 16 | 1 | 1 | 0.0625 |
| 224 | *Smarca2* | 88 | 65 | 8 | 4 | 0.061538462 |
| 241 | *Pou5f1* | 1000 | 723 | 76 | 44 | 0.060857538 |
| 350 | *Jmjd6* | 24 | 17 | 1 | 1 | 0.058823529 |
| 288 | *Carhsp1* | 83 | 52 | 11 | 3 | 0.057692308 |
| 190 | *Dnajc2* | 36 | 18 | 4 | 1 | 0.055555556 |
| 124 | *Hmgb3* | 24 | 19 | 2 | 1 | 0.052631579 |
| 305 | *Atad2* | 38 | 24 | 1 | 1 | 0.041666667 |
| 374 | *Ybx3* | 33 | 28 | 2 | 1 | 0.035714286 |
| 392 | *Lin28a* | 101 | 86 | 4 | 2 | 0.023255814 |
| 348 | *Trps1* | 63 | 56 | 1 | 1 | 0.017857143 |




**Heatmaps of ZGA-like gene expression after CRISPRa and cDNA overexpression of different screen candidates, analysed by bulk RNA-sequencing**

Heatmaps showing transcript normalised expression, scaled per transcript, of the MERVL/ZSCAN4-driven network described in Eckersley-Maslin et al. 2016, in bulk RNA-sequencing libraries after CRISPRa (top panel) and cDNA overexpression (bottom panel) of the indicated genes. Controls are two different non-targeting sgRNAs (NT1 and NT2) for CRISPRa and GFP only overexpression for cDNA. The sgRNAs used for CRISPRa are shown and sequences are available in Appendix C.

# Appendix K Western-blot analysis of *Dppa2* knock-out mESCs




**Western-blots of DPPA2, ZSCAN4, OCT4 and NANOG in *Dppa2* knock-out mESCs**

Western-blots showing DPPA2 (top left panel), ZSCAN4 (top right panel), OCT4 (bottom left panel) and NANOG (bottom right panel) protein levels in clonal wild-type and *Dppa2* knock-out mESCs. HSP90 antibody was used as a loading control.